Pipeline failure prediction in water distribution networks using weather conditions as explanatory factors

Konstantinos Kakoudakis, Raziyeh Farmani and David Butler

ABSTRACT

This paper examines the impact of weather conditions on pipe failure in water distribution networks using artificial neural network (ANN) and evolutionary polynomial regression (EPR). A number of weather-related factors over 4 consecutive days are the input of the binary ANN model while the output is the occurrence or not of at least a failure during the following 2 days. The model is able to correctly distinguish the majority (87%) of the days with failure(s). The EPR is employed to predict the annual number of failures. Initially, the network is divided into six clusters based on pipe diameter and age. The last year of the monitoring period is used for testing while the remaining years since the beginning are retained for model development. An EPR model is developed for each cluster based on the relevant training data. The results indicate a strong relationship between the annual number of failures and frequency and intensity of low temperatures. The outputs from the EPR models are used to calculate the failures of the homogenous groups within each cluster proportionally to their length.

Key words | artificial neural network, evolutionary polynomial regression, pipe failure predictions, rehabilitation, water distribution networks, weather factors

Konstantinos Kakoudakis (corresponding author)
Raziyeh Farmani
David Butler
College of Engineering, Mathematics and Physical Sciences,
University of Exeter,
Exeter EX4 4QF,
UK
E-mail: kk337@exeter.ac.uk

INTRODUCTION

The optimal management strategy for a water distribution network (WDN) balances issues of water safety, reliability, quality and quantity while exploiting the full extent of the useful life of pipes and reducing long-term costs through proactive management (Kleiner & Rajani 2001; Clair & Sinha 2012). In order to enhance this strategy, the use of predictive models is fundamental since they provide insights into the relationships between pipe failure and all the factors influencing it. These factors can be split into pipe-intrinsic, operational and environmental. Environmental and pipe-intrinsic factors can be further divided into static and dynamic (time-dependent), while the operational factors are inherently dynamic. The pipe-intrinsic factors such as the pipe material, diameter, length, age and the operational factors such as pressure, previous number of failures have been examined in several studies (e.g., Kleiner & Rajani 2001; Clair & Sinha 2012; Nishiyama & Filion 2013).

A few approaches have examined the impact of environmental factors on pipe failure trend in Canada and northern USA (Kleiner & Rajani 2002; Rajani et al. 2012; Laucelli et al. 2014), Australia (Gould et al. 2011), Netherlands (Wols & van Thienen 2013) and Austria (Fuchs-Hanusch et al. 2013).

Gould et al. (2011) conducted a statistical analysis to examine the impact of weather factors on the pipe failure of various material, diameter and soil type groups. The focus of the analysis was to relate the variation in the monthly failure rate with the dynamic weather factors. Wols & van Thienen (2013) used a linear regression analysis to ascertain the relationship between weather data and pipe failure. The analysis was conducted separately for different cohorts, depending on the type of pipe material, year of installation and diameter for a 2-month interval. Fuchs-Hanusch et al. (2013) examined the correlation between failure frequencies and climatic indicators. The winter and

summer failure frequencies were examined separately. Rajani *et al.* (2012) used a non-homogenous Poisson-based pipe deterioration model to examine the impact of air temperature-based and water temperature-based covariates on breaks of homogenous groups of pipes with respect to pipe material, age and diameter. They examined a number of (non-overlapping) time steps lasting from 5 up to 90 days, concluding that the best time step for data aggregation is 30 days. The proposed model was not validated on a test dataset since the analysis merely aimed at ascertaining the impact of temperature-based covariates on failure trends rather than using them for predictions. Laucelli *et al.* (2014) investigated the relationship between climate data and pipe bursts of 150 mm cast iron (CI) pipes using the evolutionary polynomial regression (EPR). They examined three non-overlapping time steps (5, 15 and 30 days) and concluded that the 30 days' time step provides the most accurate results. The analysis was conducted separately for the warm and cold seasons.

The failure frequency in a WDN is not constant due to the inherent nature of some of the factors affecting it. Hence, this paper examines the relationship between the annual number of pipe failures and time-dependent weather factors. The proposed approach does not require the distinction between cold and warm seasons to be made. The approach for making annual predictions is complementary to a previous study (Kakoudakis *et al.* 2017), which calculated the average failure rate for a specific period using pipe-related factors as explanatory variables. The failure frequency is the cumulative effects of several factors on the pipes, therefore the results of the two approaches are combined. Furthermore, a method is proposed to identify more vulnerable regions of the network and visualise them on a map.

The occurrence of pipe failures requires the fast response of the network's operators. The water companies aim to respond as soon as possible after a burst is reported to minimise the amount of lost water and the customer dissatisfaction that might result in need for compensation. The response time depends on several factors including, among others, the availability of human resources and the ability to predict time intervals with an above the normal failure frequency. Previously developed approaches have resulted in relationships with low accuracy (i.e., Rajani *et al.* 2012)

for short-term predictions. This paper proposes a method to predict the occurrence of pipe failure(s) on a short period without requiring knowledge of the weather). In addition, the weather-related factors are ranked based on their importance for predictions.

It should be noted here that the annual predictions can be used in conjunction with long-term predictions for pipe maintenance/rehabilitation/replacement scheduling while the short-term predictions are strictly for operational use. Furthermore, the results for the annual predictive models are on a cluster level while the short-term predictions refer to the entire examined network (due to the small number of failures in some clusters) and do not associate the failure occurrence with specific pipes.

The remainder of the paper is organised as follows: First the proposed methodologies are explained. The section 'EPR' provides a description of the software used. Then the process to evaluate the accuracy of the proposed methodology is explained. The main features of the case study are provided next. This is followed by discussion of the results and the evaluation of their accuracy. Finally, the last section highlights the most important conclusions.

METHODOLOGY

Annual pipe failure prediction

The annual failure rate is an important performance indicator for assessing the overall structural condition of a WDN (Fuchs-Hanusch *et al.* 2013) and, therefore, models that predict it are of significant interest. This paper presents a method for predicting the next year's number of failures considering weather conditions as explanatory variables. Furthermore, outputs from the models are used to calculate the failure rates of individual pipes in order to identify regions of the network that are more prone to failure. The methodology consists of the following steps:

1. First, individual pipes are *aggregated* into homogenous groups based on their diameter, the age and the soil type assuming that pipes that share the same characteristics are expected to have a similar failure rate (Kleiner & Rajani 2012). Soil type is used as an *aggregation*

criterion because soil properties have been associated with the corrosion of metallic pipes (Sadiq *et al.* 2004), which has been identified as a dominant factor contributing to their failure (Makar 2000). The original dataset containing a large number of individual pipes is converted to a new dataset containing homogenous groups of pipes.

2. The created homogenous groups of pipes are allocated into six clusters using their attributes of diameter and age based on the findings of a previous analysis (Kakoudakis *et al.* 2017) that demonstrated how splitting the network into six clusters using the K-means clustering method could result in more accurate predictions. Instead of using a single model for making predictions for all the homogenous groups, six separate predictive models are developed.

3. For each cluster the annual number of failures is equal to the sum of failures of the homogenous groups within. The candidate weather-related explanatory variables are: average minimum air temperature (Equation (1)), average maximum air temperature (Equation (2)), average soil temperature (Equation (3)) and freezing index which is calculated only for the days below a predefined threshold (Equation (4)):

$$AveT_{min} = \frac{\sum_{j=1}^{m} T_{minimum}^{j}}{m} \tag{1}$$

$$AveT_{max} = \frac{\sum_{j=1}^{m} T_{maximum}^{j}}{m} \tag{2}$$

$$AveST = \frac{\sum_{j=1}^{m} ST^{j}}{m} \tag{3}$$

$$FI = \sum_{j=1}^{m} \left(\theta - T_{minimum}^{j} \right) \tag{4}$$

where m is the number of days in the time step (i.e., 365 days), $T_{minimum}^{j}$ is the minimum daily temperature of day j, $T_{maximum}^{j}$ is the maximum daily temperature of day j, ST^{j} is the average daily soil temperature of day j and θ is the predefined air temperature threshold.

4. The freezing index (Equation (4)) is defined as the cumulative minimum daily temperature below a specified air temperature threshold and is considered as a surrogate for the severity of extreme air temperatures within a time step (Kleiner & Rajani 2002). The cross-correlation function in MATLAB (®R2014b) is applied to measure the similarity between the candidate thresholds and the number of failures. The thresholds examined range between −2 °C and 4 °C with a step of 1 °C. This process is repeated separately for each cluster. The threshold that provided the highest similarity (highest values of cross-correlation) was selected for data aggregation.

5. The EPR models are selected with respect to their goodness of fit, the minimisation of model's polynomial terms and the possibility to describe the physical phenomenon. The predicted numbers are used to calculate the number of failures for the homogenous groups within each cluster proportionally to their length.

6. Then, it is assumed that all the pipes within those homogenous groups have the same failure rate for this specific year. These values are used in combination with the complementary approach (Kakoudakis *et al.* 2017) to calculate the final failure rate for this year.

Daily prediction of the occurrence of pipe failures

The proposed method aims to predict the occurrence of failure and consists of the following steps:

1. Define the inputs and the output of the model. The inputs of the ANN model are: the minimum air temperature, the maximum air temperature, the mean air temperature, the soil temperature, the freezing index and the number of failures for a number of consecutive days while the targeted output of the model is 1 if there is at least a pipe failure in the following few days and 0 if not. The temperature variation can occur relatively quickly whereas the potential pipe failure because of that might take longer (Rajani *et al.* 2012). Therefore, different combinations of number of days are examined in order to obtain the models with the highest accuracy. Exhaustive trials were conducted leading to the conclusion that the use of 4 consecutive days as input and the following 2 days as output results in the highest accuracy. The first input is the set of variables for the first 4 days and the output is the occurrence of failure(s) in the 5th and 6th day. Respectively, the second input is the set of variables

from the 2nd up to the 5th day, while the output is the failure in the 6th and 7th day.

2. The inputs and the outputs are divided into two parts, for training (70%) and test (30%). The ANN model is built relying only on the training data.

3. The actual output of the model is not an integer number; therefore the optimal threshold for converging to 1 (failure) or 0 (non-failure) has to be identified. The selection of the optimal threshold entails three steps:

   3a Initially, a set of candidate thresholds covering the entire range between the model's minimum and maximum responses for the test data is defined. Then, the model's actual outputs are rounded (to 1 or 0, respectively) for all the values of candidate thresholds.

   3b The true positive rate (*TPR*) and the false positive rate (*FPR*) are calculated for all the candidate thresholds. This iterative process provides a set of *TPR/FPR* pairs which are used to plot the receiver operating characteristic (ROC) curve. Each point on the ROC plot (Figure 1) represents a specific *TPR/FPR* pair. A model with perfect discrimination has a ROC curve that passes through the upper left corner (optimal point) (Zweig & Campbell 1993). On the contrary, the closer the curve comes to the 45° diagonal of the ROC space, the less accurate the model is. Therefore, the most accurate curve is the C and the least accurate the A.
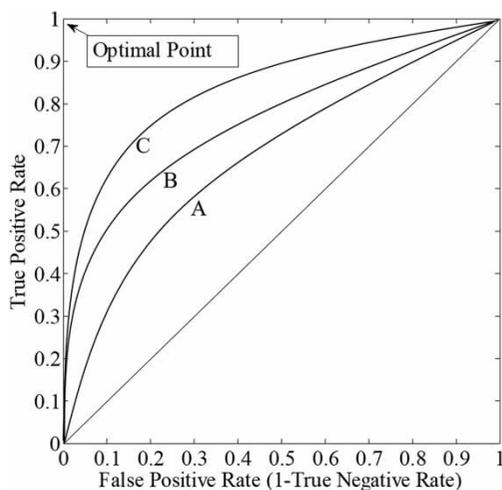

Figure 1 │ The general form of receiver operating curve.

3c The Euclidian distance (distance between each point on the curve and the optimal point) is calculated as follows:

$$Euclidian\,distance = \sqrt{(1 - TPR)^2 + (FPR)^2} \qquad (5)$$

The threshold with the minimum Euclidian distance is selected since it provides the most accurate results by minimising the false positive rate and maximising the *TPR*.

4. At the last stage, the influence of the inputs on the model's response is assessed. The analysis is performed using the following equation (Duncan *et al.* 2013):

$$W_{io} = W_1 * W_2 \qquad (6)$$

where $W_{io} =$ input-to-output influence vector; $W_1 =$ ANN hidden layer weight matrix; $W_2 =$ ANN output layer weight vector. Thus, W_{io} has dimensions of $N_{in}*N_{out}$ where N_{in} is the number of inputs and $N_{out} = 1$ is the number of output neurons.

EVOLUTIONARY POLYNOMIAL REGRESSION

EPR (Giustolisi & Savic 2006) is a data-driven method which combines numerical and symbolic regression. The implementation of EPR returns a predefined number of models on a Pareto optimal front which is a trade-off between accuracy and parsimony. The accuracy criterion aims to maximise the model fit to the observed data and the parsimony to minimise the number of explanatory variables and/or polynomial terms in the model. The role of the parsimony rule is to prevent over-fitting of the model to data and thus capture the true underlying general phenomena (Laucelli *et al.* 2014). The user selects the generalised model structure and EPR employs a multi-objective search strategy to estimate the unknown parameters. The model structure selected here for analysis of pipe failure is (Giustolisi & Savic 2006):

$$Y = \sum_{j=1}^{\kappa} a_j \left((X_1)^{E_{1j}} \dots (X_i)^{E_{ij}} \right) + a_0 \qquad (7)$$

where Y = predicted number of pipe failures; a_j and a_0 = the constant coefficients; X_i = is the explanatory variable i, E_{ij} = the matrix of unknown exponents and k is the maximum number of polynomial terms.

The candidate exponent values (E_{ij}) in Equation (7) were −2, −1, −0.5, 0, 0.5, 1 and 2 describing potential square, linear or square root exponents for explanatory variables of the EPR model. The positive and negative values were considered to describe potential direct and inverse relationship between the inputs and the output of the model while the value 0 was chosen to deselect input candidates without impact on the output. The maximum number of polynomial terms (k) was set to 1 excluding the constant term (a_0) to ensure the best fit without unnecessary complexity, as the addition of new terms fit mostly random noise in the raw data rather than explain the underlying phenomenon. The least square (LS) parameter was constrained to search for positive polynomial coefficient values only (i.e., $a_j > 0$) because negative polynomial coefficients usually try to balance positive terms providing a better description of the noise (Giustolisi *et al.* 2007).

MODEL PERFORMANCE ASSESSMENT

The performance indicator used to evaluate the accuracy of the EPR models is the coefficient of determination (R^2) as a measure for correlation between predictions and observations. The mathematical relationship is expressed as follows (Moriasi *et al.* 2007):

$$R^2 = \frac{\left(\sum_{i=1}^{n} (y_{p,i} - \bar{y}_p)(y_{o,i} - \bar{y}_o) \right)^2}{\sum_{i=1}^{n} (y_{p,i} - \bar{y}_p)^2 \sum_{i=1}^{n} (y_{o,i} - \bar{y}_o)^2} \tag{8}$$

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 n = the number of test data samples.

The performance of the binary model is assessed using the *TPR* (Equation (9)) and true negative rate (Equation (10)). *TPR* measures the proportion of correctly identified positives while *TNR* measures the proportion of correctly identified negatives, respectively. The mathematical

expressions of *TPR* and *TNR* are defined as (Kohavi & Provost 1998):

$$True\,Positive\,Rate = \frac{True\,Positives}{True\,positives + False\,negatives} \tag{9}$$

$$True\,Negative\,Rate = \frac{True\,Negatives}{True\,negatives + False\,positives} \tag{10}$$

CASE STUDY

The proposed methodology is demonstrated in a case study which is part of a WDN of a UK city. The database contains pipe failure data between 1st January 2003 and 31st December 2013. Preliminary analysis showed that CI pipes which constitute 78% of the network's total length have the highest pipe failure rate (expressed in number of failures/km/year) which is 0.264 compared to other pipe material types which are 0.194 for asbestos cement (AC) pipes, 0.071 for ductile iron (DI) pipes, 0.030 for polyethylene (PE) pipes and 0.113 for polyvinyl chloride (PVC) pipes. Hence, only CI pipes are considered in this paper for construction of the predictive models. Table 1 shows the main features of the examined dataset.

Daily climate data for the case study were obtained from the British Atmospheric Data Centre and consisted of the minimum air temperature, the maximum air temperature and the soil temperature on a daily basis in °C. To avoid negative values, all temperatures are converted to Fahrenheit.

Preliminary analysis of the data showed (Figure 2) that the majority of the failures occur during the coldest

Table 1 │ The main features of the cast iron pipes in the case study

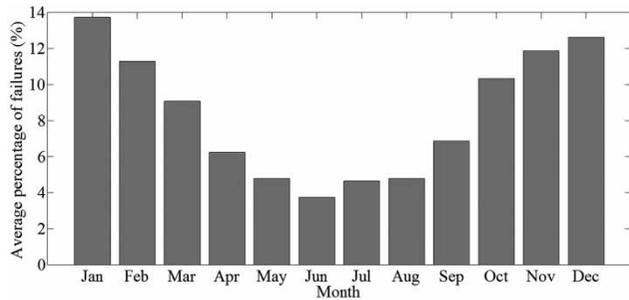| Feature	Value/range
Installation year	1,865–1,995
Diameter range	75–300 mm
Total length	647.00 km
Number of pipes	18,872
Number of failed pipes	1,089
Number of failures	1,442

Figure 2 | Average percentage of pipe failures per month.

months. Therefore, in the development of the models, a particular emphasis is given to the factors that describe the severity of the cold period (i.e., freezing index).

RESULTS AND DISCUSSION

Results of the annual predictions approach

Following the procedure described above for the data preparation, grouping of individual pipe failure data resulted in 148 homogenous groups for developing the EPR models. Those homogenous groups were split into six clusters as shown in Figure 3. The dataset created was split into two parts for model development and validation, respectively. The last year (i.e., 2013) of the monitoring period was used for validation purposes.

The implementation of the proposed methodology resulted in six EPR models, each corresponding to the training data of the relevant cluster. The selected threshold for the freezing index is 0 °C degrees because it provided the highest correlation in the preliminary analysis. Table 2 lists the associated models and the coefficient of determination for the training dataset.
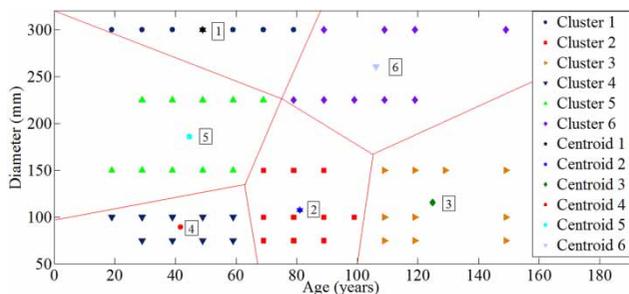

Figure 3 | Input data clustering with the six clusters and the corresponding centroids.

**Table 2** | Obtained models for the six clusters

Six-clustered EPR	R^2
Cluster 1: $C_1 = 0.001(FI^{0.5}) + 0.46$	0.80
Cluster 2: $C_2 = 2.513(FI^{0.5}) + 8.52$	0.78
Cluster 3: $C_3 = 1.834(FI^{0.5}) + 5.21$	0.84
Cluster 4: $C_4 = 3.425(FI^{0.5}) + 12.76$	0.93
Cluster 5: $C_5 = 1.143(FI^{0.5}) + 4.05$	0.75
Cluster 6: $C_6 = 0.01(FI^{0.5}) + 1.32$	0.86

The relationship between the number of failures and the freezing index is a direct indication that lower temperatures and consequently higher values of the freezing index cause an increase in the number of failures. The addition of more candidate explanatory variables (e.g., minimum air temperature, maximum air temperature, soil temperature) did not increase the model's accuracy and therefore they were not selected.

Figure 4 shows the predictions vs the observations for all the clusters with the test dataset (2013). The developed models are very accurate in predicting the number of failures for clusters 1 and 6, which have the lowest number of failures. The absolute difference between observations and predictions for clusters 1 and 6 tends to zero whereas it varies between 3 and 6.5 for the rest of the clusters. The lowest error is achieved for clusters 1 and 6, which have the lowest failure rate.

The predicted number of failures was used to calculate the number of failures for all the homogenous groups (i.e., diameter, age, soil type) within the clusters proportionally to their length. Then, it was assumed that all the individual pipes within the homogenous groups share the same failure rate. The individual pipe failure rates were classified using the Jenks Natural Breaks (Jenks 1963) method into five ranges as 'very low' [0–0.091], 'low' (0.091–0.236], 'medium' (0.236–0.472], 'high' (0.472–0.75] and 'very high' [greater than 0.751] as shown in Figures 5 and 6 (observations and predictions respectively).

The accuracy obtained in allocating the individual pipes in ranges is 46%, 73%, 78%, 87% and 76% for the 'very low', 'low', 'medium', 'high' and 'very high' failure rates, respectively, when only weather-related factors are used. The predictions have a high accuracy for the majority of the
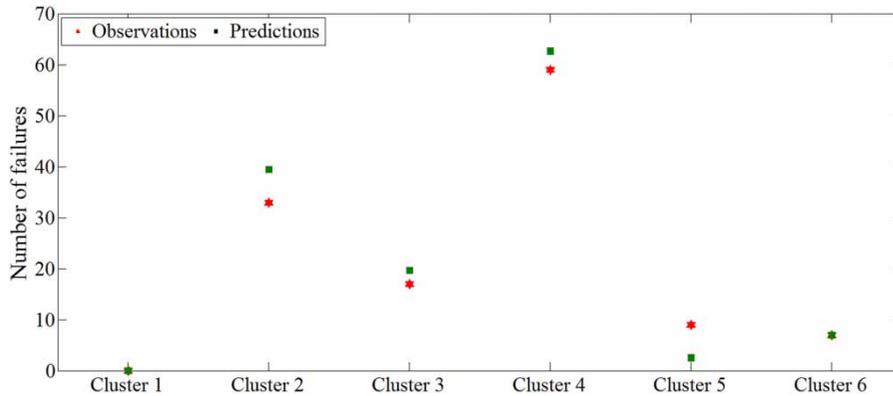
Figure 4 | Observations and predictions for the six clusters.
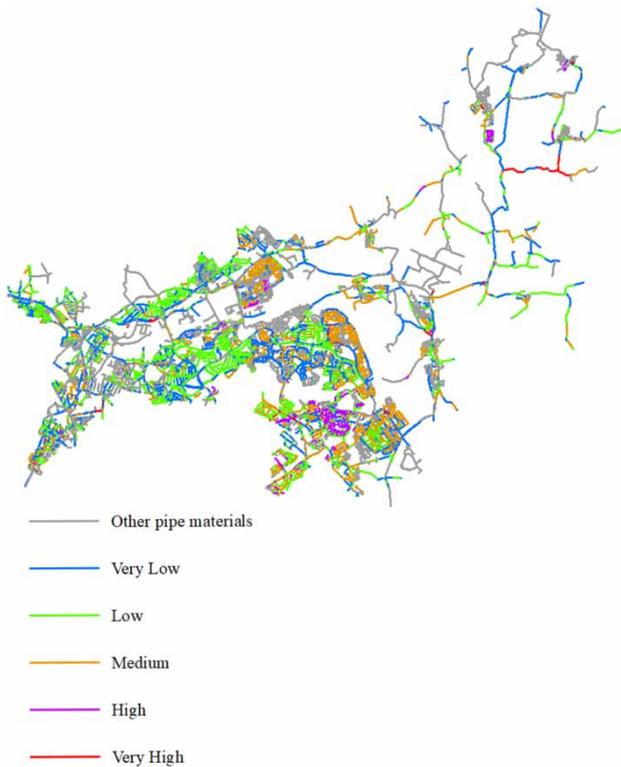
Figure 5 | Observed pipe failure rates in 2013.
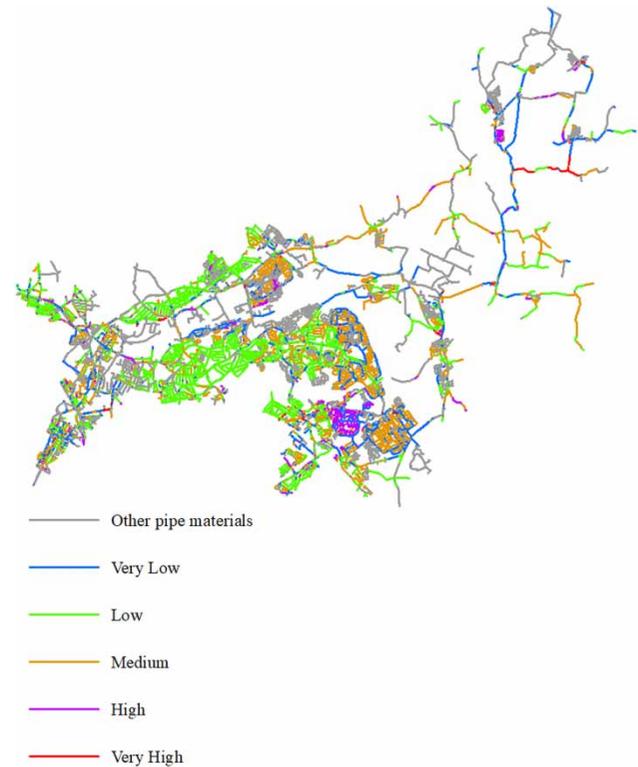
Figure 6 | Predicted pipe failure rates in 2013.

failure ranges ('low' to 'very high'). The lowest accuracy is achieved for the pipes with a 'very low' observed failure rate. This low accuracy can be attributed to the fact that a number of homogenous groups of pipes have experienced zero number of failures. The predicted failures for each cluster are distributed to the homogenous groups proportionally to their length value leading to a slight overestimation for those groups.

Results of combined weather and pipe-intrinsic factors-based approach

The outputs of the proposed method are used in conjunction with the results of an approach which calculated the average failure rate for the entire monitoring period using pipe-intrinsic factors as explanatory variables (Kakoudakis *et al.* 2017) as shown in Figure 7. The final failure rate
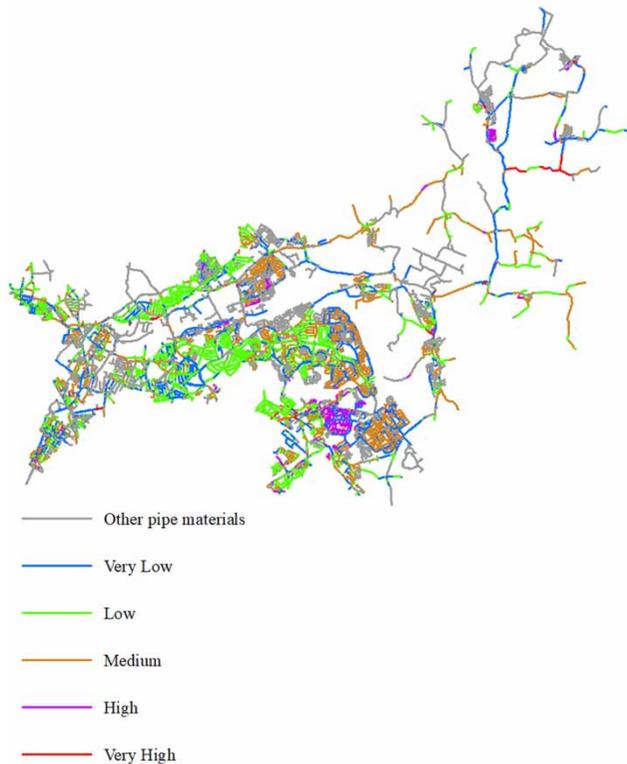
Figure 7 │ Predicted pipe failure rates in 2013 combining the two approaches.

of the individual pipes is the combination of the two values.

The examined WDN consists of a big number of individual pipes and the improvement achievement is highlighted in Figure 8. Figure 8 compares the accuracy of the predictions when only environmental variables are used and when they are combined with the physical variables. The inclusion of the physical factors increased the accuracy of the predictions for the majority of the ranges. The highest improvement is observed for the 'very low' range which shifted to 69%.
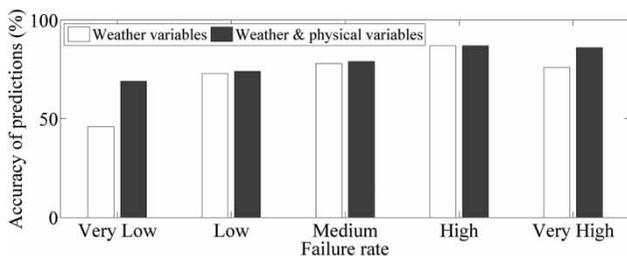
Results for the short-term predictions

Following the approach described in the methodology section, the data preparation resulted in 3,653 data samples, 56.20% of which correspond to cases without failure(s) and the remaining 43.80% to failure(s). The model's responses were compared to a set of threshold values and the generated pairs of *TPR/FPR* were used to plot the ROC curve (Figure 9). The selected threshold with the lowest Euclidean distance from the optimal point is 0.538. As shown in Figure 9, the majority of the non-failures are correctly identified (the *FPR* is 0.87) as such while a similar conclusion can be derived for the failures despite the lower accuracy (the *TPR* is 0.72). The value of *AUC* which is used as a measurement of the model's performance is 0.814, indicating that the model has a good accuracy.

The influence of the inputs on the model's response is assessed and they are ranked to identify the most influential. The result of the analysis input's influence on the model's responses is a column matrix with the weight of all the inputs (Table 3). The *FI* is shown to be the most influential factor. This observation is linked to the fact that the majority of the failures occur in the coldest months (also shown in Figure 2) when pipes are subject to frost actions, which is a cause for axial causes on them (Rajani *et al.* 1996). The frost imposes additional load on the buried pipes and is influenced by frost penetration, trench width, soil type, soil stiffness, frost heave of trench fill and side fill as well as
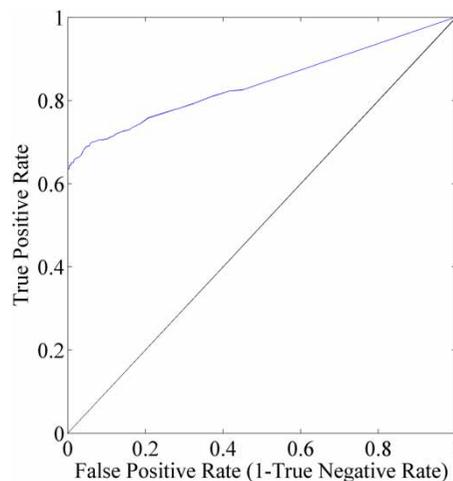

Figure 8 │ Percentage of predictions accuracy in different ranges.

Figure 9 │ Receiver operating curve of the binary ANN model.

Table 3 | The weight of the weather-related inputs on the response

Input	Weight
FI_{t-2}	15.7495
FI_{t-1}	13.0278
FI_{t-3}	12.3511
$MinT_{t-2}$	−11.4116
$MinT_{t-4}$	−11.0194
Max_{t-2}	−10.6109
$Soil_{t-2}$	−10.261
$MinT_{t-1}$	−9.5467
$Soil_{t-4}$	−9.5358
$MaxT_{t-4}$	−9.4932
$Soil_{t-3}$	−8.6878
$MaxT_{t-1}$	−8.3463
$Soil_{t-1}$	−7.6562
$MaxT_{t-3}$	−7.5278
$MinT_{t-3}$	−7.3292
FI_{t-4}	6.317

the interaction at the trench backfill–side fill interface (Rajani & Zhan 1996). The negative values indicate a reverse relationship between these variables and the occurrence of pipe failure(s).

CONCLUSIONS

This paper presents a method to predict the occurrence of pipe failure and its annual variation due to weather factors. Only CI pipes were considered due to their highest failure rate in the network. However, it can be applied to other pipe materials as well. For the annual predictions the individual pipes were allocated into homogenous groups. The created homogenous groups were then split into a predefined number of clusters and an individual EPR model was developed for each cluster using weather conditions as explanatory variables. The *FI* was selected as the most influential variable by the models. The mathematical relationship obtained between the number of failures and the *FI* is a direct indication that lower temperatures and consequently higher values of the *FI* cause an increase in the number of failures. The outputs of the proposed method were used in

conjunction with a previous approach which calculated the average failure rate of the entire monitoring period using pipe-intrinsic explanatory variables in order to improve the quality of predictions. The final failure rate was calculated as the average failure rate of the two approaches which resulted in more accurate predictions. The highest improvement was achieved for the pipes with a 'very low' observed failure rate.

The method for predicting the occurrence of failure(s) was implemented using an ANN binary model and was shown to be able to distinguish between the days with and without failure(s). The influence of inputs on the models' output responses was assessed showing that low temperatures have a strong influence. This approach can be used operationally to alert water utilities to manage pipe failures reducing potential water loss, associated costs and service disruption to consumers. Further research should be done to associate the short-term prediction of failure(s) with specific pipes.

REFERENCES

Clair St, A. M. & 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.

Duncan, A., Tyrrell, D., Smart, N., Keedwell, E., Djordjevic, S. & Savic, D. 2013 Comparison of machine learning classifier models for bathing water quality exceedances in UK. In: *Proceedings of the 35th International Association for Hydro-Environment Engineering and Research (IAHR 2013) Congress*, Chengdu, China.

Fuchs-Hanusch, D., Friedl, F., Scheucher, R., Kogseder, B. & Muschalla, D. 2013 Effect of seasonal climatic variance on water main failure frequencies in moderate climate regions. *Water Science & Technology: Water Supply* **13** (2), 435–446.

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

Giustolisi, O., Doglioni, A., Savic, D. A. & Webb, B. W. 2007 A multi-model approach to analysis of environmental phenomena. *Environmental Modelling and Software* **22** (5), 674–682.

Gould, S. J. F., Boulaire, F. A., Burn, S., Zhao, X. L. & Kodikara, J. K. 2011 Seasonal factors influencing the failure of buried water reticulation pipes. *Water Science & Technology* **63** (11), 2692–2699.

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

Kakoudakis, K., Behzadian, K., Farmani, R. & Butler, D. 2017 Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with *K*-means clustering. *Urban Water Journal* **14** (7), 737–742.

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

Kleiner, Y. & Rajani, B. 2002 Forecasting variations and trends in water-main breaks. *Journal of Infrastructure Systems* **8** (4), 122–131.

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

Kohavi, R. & Provost, F. 1998 Glossary of terms. *Machine Learning* **30** (2–3), 271–274.

Laucelli, D., Rajani, B., Kleiner, Y. & Giustolisi, O. 2014 Study on relationships between climate-related covariates and pipe bursts using evolutionary-based modelling. *Journal of Hydroinformatics* **16** (4), 743–757.

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. & Veith, T. L. 2007 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* **50** (3), 885–900.

Nishiyama, M. & Filion, Y. 2013 Review of statistical water main break prediction models: 2002–2012. *Canadian Journal of Civil Engineering* **40** (10), 972–979.

Rajani, B. & Zhan, C. 1996 On the estimation of frost loads. *Canadian Geotechnical Journal* **33** (4), 629–641.

Rajani, B., Zhan, C. & Kuraoka, S. 1996 Pipe–soil interaction analysis of jointed water mains. *Canadian Geotechnical Journal* **33** (3), 393–404.

Rajani, B., Kleiner, Y. & Sink, J. 2012 Exploration of the relationship between water main breaks and temperature covariates. *Urban Water Journal* **9** (2), 67–84.

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

Wols, B. A. & van Thienen, P. 2013 Impact of weather conditions on pipe failure: a statistical analysis. *Journal of Water Supply: Research and Technology* **63** (3), 212–223.

Zweig, M. H. & Campbell, G. 1993 Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39** (4), 561–577.