Development of pipe deterioration models for water distribution systems using EPR

L. Berardi, O. Giustolisi, Z. Kapelan and D. A. Savic

ABSTRACT

The economic and social costs of pipe failures in water and wastewater systems are increasing, putting pressure on utility managers to develop annual replacement plans for critical pipes that balance investment with expected benefits in a risk-based management context. In addition to the need for a strategy for solving such a multi-objective problem, analysts and water system managers need reliable and robust failure models for assessing network performance. In particular, they are interested in assessing a conduit's propensity to fail and how to assign criticality to an individual pipe segment. In this paper, pipe deterioration is modelled using Evolutionary Polynomial Regression. This data-driven technique yields symbolic formulae that are intuitive and easily understandable by practitioners. The case study involves a water quality zone within a distribution system and entails the collection of historical data to develop network performance indicators. Finally, an approach for incorporating such indicators into a decision support system for pipe rehabilitation/replacement planning is introduced and articulated.

Key words | data-driven modelling, evolutionary polynomial regression, failure analysis, performance indicators, water systems

L. Berardi (corresponding author)
O. Giustolisi
Hydroinformatics Group,
Technical University of Bari,
via Orabona 4 I-70125 Bari,
Italy
E-mail: *l.berardi@poliba.it*

Z. Kapelan
D. A. Savic
School of Engineering and Computer Science,
Harrison Building,
University of Exeter,
North Park Road, Exeter EX4 4QF,
UK

ABBREVIATION AND NOTATION

Ap	Pipe age
A, A_{class}	Equivalent age of the pipe *class*
$A_{0,class}$	Equivalent age of the pipe *class* at the end of the monitoring period.
a_j	jth constant value in polynomial expressions
Brt	Bursts recorded (total) for the pipe *class*
Brp_i	Bursts recorded for the i-th pipe
BR, BR_{class}	Bursts predicted for the pipe *class*
BR_i	Burst predicted for the ith pipe
CoD	Coefficient of determination
d_i	Damage subsequent to failure of pipe i
Dp	Pipe's nominal diameter
D, D_{class}	Equivalent diameter of the pipe *class*
DSS	Decision support system
EPR	Evolutionary polynomial regression
ES	Matrix of exponents of EPR input variables
f, g	Functions selected by user in the EPR model structure
h	Planning horizon for burst predictions
L, L_{class}	Total length of the pipe *class*
Lp	Pipe length
LS	Least squares
m	Number of polynomial terms of the expressions returned by EPR
MOGA	Multi-objective genetic algorithm
n	Number of samples of observed data
N, N_{class}	Number of pipes in the pipe *class*
Pp	Number of properties supplied by a pipe
P, P_{class}	Number of properties (total) supplied by the pipe *class*
t	time variable
T	monitoring time period
WQZ	Water quality zone
SSE	Sum of squared errors
$Vp_{i,s}$	Value of the sth aggregate variable selected in the EPR model for pipe i

doi: 10.2166/hydro.2008.012

$V_{class,s}$	Value of the sth aggregate variable selected in the EPR model for *class*
$\mathbf{X}_k$	kth candidate input variable
Xi	Number of input variables selected in the ERP model
$\mathbf{Y}$	Vector of target values
$\hat{y}$	Value returned by the model
$y_{\exp}$	Observed value
α	Exponent of variable A in the EPR model
γ	Exponent of variable L in the EPR model
δ	Exponent of variable D in the EPR model
ρ	Exponent of variable P in the EPR model
μ	Exponent of variable N in the EPR model
λ_i^{EPR}	Failure rate of pipe i according to EPR model
λ_i^{R}	Failure rate of pipe i according to recorded bursts

INTRODUCTION

Pipe bursts are a regular occurrence in water distribution systems. Bursts commonly occur when the residual strength of a deteriorated main becomes inadequate to resist the force imparted on it (Skipworth *et al.* 2002). From a terminology point of view pipe bursts are commonly referred to also as *breaks* or *failures* and are linked to *leaks* when losses in water distribution networks are analysed (Farley & Trow 2003). The deterioration of pipes may be classified into two categories (Kleiner & Rajani 2001): (1) *structural* deterioration, which diminishes the pipe's structural resilience and its ability to withstand the various types of stresses imposed upon it; (2) *functional* deterioration of inner surface of the pipe resulting in diminished hydraulic capacity and degradation of water quality.

The consequence of pipe failures is not only an economic burden (repair and other costs), but it can also have significant social (e.g. service interruptions, traffic delays, etc.) and environmental (e.g. lost water and energy) impacts. A number of research projects have been recently undertaken with the goal of developing a Decision Support System for optimal asset management of water and wastewater systems (LeGauffre *et al.* 2002; Skipworth *et al.* 2002; Savic *et al.* 2005). An integral part of these projects is the selection of Performance Indicators (PIs) and their integration into the decision-making process (McDonald & Zhao 2001; Shepherd *et al.* 2004, Giustolisi *et al.* 2006*a*). IWA best practice manuals (Alegre *et al.* 2000; Matos *et al.* 2003) are often taken as a reference point for defining and selecting relevant PIs which are typically derived by modelling hydraulic behaviour and asset performance. Both types of models are based on the analysis of existing water company data related to physical infrastructure and on the historical records of associated failure events.

It is worth saying that more often than not the research efforts in introducing new and more sophisticated data analysis techniques become futile as data are either not available or scarce in terms of quality (e.g. because of poor collection methodologies) and quantity (e.g. short recording period). This evidence highlights a serious responsibility of water companies and municipalities in maintaining adequate data collection levels.

Among the different studies carried out on deriving structural deterioration models, a preliminary distinction has to be made between physically based approaches and statistical methods (Kleiner & Rajani 2001). The former aim at describing the physical mechanisms underlying pipe failure and require data that is costly or impossible to obtain. The latter can be applied with variable input data quality and may be useful even when only limited data is available. For water distribution pipes, statistical models provide a cost-effective means of analysis.

Water mains deterioration has traditionally been studied as a steady monotonic process affected by time-varying "noise" (Kleiner & Rajani 2002). Time-dependent factors can be random, cyclical (i.e. environmental conditions) or variable (i.e. operational factors), often resulting in a masking effect of the underlying ageing patterns, especially in small datasets. The effectiveness of analysing these factors depends primarily on the accuracy of forecasting the time-related phenomena (e.g. weather conditions) and on the planning horizon adopted (i.e. short-term vs. long-term rehabilitation).

Hitherto, the majority of statistical models developed consider pipe age as the most important variable describing the time dependence of pipe breakage. Exponential (Shamir & Howard 1979; Walski & Pelliccia 1982) and time-powered

models (Mavin 1996; Kleiner & Rajani 2001) have been used to determine the optimal timing of pipe replacement, with both approaches exhibiting comparable accuracy and performance (Mavin 1996).

Two important observations made by a number of researchers are: (1) age is not the only governing parameter of pipe breaks (Walski & Pelliccia 1982; Clark *et al.* 1982; Kettler & Goulter 1985) and (2) pipes often need to be aggregated into homogeneous groups in order to conduct more effective analysis (Shamir & Howard 1979; Lei & Saegrov 1998; Kleiner & Rajani 1999).

In addition to age, pipe diameter was identified early on as a key factor affecting pipe failure rates (Walski & Pelliccia 1982; Clark *et al.* 1982). In particular, a strong inverse correlation was found between pipe diameter and failure rate (Kettler & Goulter 1985), with small diameter pipes evincing higher breakage rates than their larger counterparts.

The spatial and temporal clustering of pipe failures was first done by Goulter & Kazemi (1988) and further investigated by Jacobs & Karney (1994). The main outcomes of the latter study were the definition of independent breaks (i.e. failures that occur at least 90 d after, and more than 20 m away from, the previous failure) and the observation that these breaks are uniformly distributed along the length of the water mains.

Studies examining metallic pipe behaviour (i.e. cast iron, ductile iron, etc.) have been carried out to establish the influence of pipe material on breakage rates (Kettler & Goulter 1985; Kleiner & Rajani 2002). That performed on a real network by Pelletier *et al.* (2003) revealed that a close dependence exists among pipe material, diameter and the year the pipe was laid.

The need for aggregating pipes into homogenous classes results from the small number of failures usually available for a given network, making development of a statistical model for individual pipes difficult to accomplish. Shamir & Howard (1979) were the first to suggest that data groups ought to be considered as homogeneous with respect to the causes of failure. Pipe material, diameter and age, with or without additional factors such as soil types and/or land use above the pipes, have been widely adopted as grouping criteria to emphasise their influence on failure (Herz 1996; Lei & Saegrov 1998; Le Gat & Eisenbeis 2000). Some pipe break models include such indicator variables of aggregated pipes in their formulations. This is the case for the proportional hazard (Andreou *et al.* 1987*a*, *b*), the time-dependent Poisson (Constantine *et al.* 1996) and the accelerated Weibull hazard (Le Gat & Eisenbeis 2000) models. Despite different underlying philosophies and variables considered, all of these approaches aim at describing pipe break rates with a unique expression in which all pipes share the same explanatory variables.

It is worth noting that some authors (e.g. Le Gat & Eisenbeis 2000) have included pipe length as an additional explanatory variable and have made a distinction between pipes with no failures and those with a failure history (Andreou *et al.* 1987*a*, *b*). Such a distinction is consistent with both statistical findings (Goulter & Kazemi 1988) and the description of the life cycle of a buried pipe, usually represented as a "bathtub" curve (Andreou *et al.* 1987*a*, *b*; Kleiner & Rajani 2001; Watson 2005).

Recently, Watson (2005) employed a hierarchical Bayesian model that uses both information and engineering knowledge obtained from aggregated pipes when deriving failure rate estimates for an individual pipe. This is achieved by assuming that the underlying failure rates for pipes with similar characteristics are drawn from the same prior distribution. The influence of other factors, such as pipe length, pressure, diameter, material, installation date, and soil type, is also incorporated into a proportional intensity model. Such a probabilistic approach allows for formal measurement of the uncertainty of an individual pipe's failure rate, even though it requires the elicitation of expert knowledge.

In parallel with the statistical approaches mentioned above, the complexity of water networks have led to the recent employment of data mining techniques (Fayyad *et al.* 1996) to discover patterns in pipe failures data sets (Bessler *et al.* 2002; Babovic *et al.* 2002). In particular, a novel hybrid data-driven technique, Evolutionary Polynomial Regression (EPR) (Giustolisi & Savic 2006), has been used for modelling failures in urban water systems (Berardi *et al.* 2005; Savic *et al.* 2006). The main advantages of models returned by EPR are their parsimony, the possibility of testing their physical

consistency and an intuitive way for including engineering judgement into the process of model construction and selection. Moreover, EPR aggregate models are usually accurate in describing failure occurrence in homogeneous pipe groups. All these features and the encouraging results obtained previously make EPR preferable over other modelling techniques purely based on either regressive algorithms or probabilistic approaches.

This paper describes the analysis of an asset database containing an inventory of all pipes and related bursts for a UK water distribution system. In the case study presented, information on pipe diameter, material, length, year laid, number of properties supplied and the total number of burst events recorded are available at the individual pipe level. A data organization scheme that puts data into homogeneous classes useful for the subsequent modelling phase is presented first. The application of EPR is demonstrated next and an aggregate mathematical model for pipe burst prediction is developed. A methodology to derive individual pipe structural deterioration models from aggregate EPR models is also introduced. Finally, the use of such a model in a decision-making context is outlined.

EVOLUTIONARY POLYNOMIAL REGRESSION (EPR)

In this section, a brief description of EPR methodology and features is presented (further mathematical details about EPR can be found in Giustolisi & Savic (2006) and the EPR website (see reference list)). EPR belongs to the family of Genetic Programming strategies (Koza 1992) and, according to the categorization of modelling techniques based on transparency level (Ljung 1999; Giustolisi 2004; Giustolisi & Savic 2006), it may be classified as a grey box technique. Accordingly, the approach is based on observed field data while also permitting the introduction of prior insight into the system or problem at hand. Moreover, the mathematical structures it returns are symbolic and usually parsimonious.

The EPR methodology offers two main stages: (1) search for the best model structure using an integer-coded MOGA (Multi-Objective Genetic Algorithm) (Giustolisi *et al.* 2006*b*) and (2) parameter estimation for an assumed model structure using the least squares (LS) method (Draper & Smith 1998). When performing the search for a best model structure, a generalized true and/or pseudo-polynomial model structure is assumed. The following general model structures are considered (Giustolisi & Savic 2006):

$$\text{case } 0: \quad \mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \cdot f\left((\mathbf{X}_1)^{\mathbf{ES}(j,k+1)}\right) \cdot \ldots \cdot f\left((\mathbf{X}_k)^{\mathbf{ES}(j,2k)}\right)$$

$$\text{case } 1: \quad \mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \cdot f\left((\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)}\right) \tag{1}$$

$$\text{case } 2: \quad \mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \cdot f\left((\mathbf{X}_1)^{\mathbf{ES}(j,k+1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,2k)}\right)$$

$$\text{case } 3: \quad \mathbf{Y} = g\left(a_0 + \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)}\right)$$

where $\mathbf{X}_k$ is the kth explanatory variable, **ES** is the matrix of unknown exponents (coded as integers in the MOGA, representing ordinal numbers of optional exponent values, as defined by the user), f and g are functions selected by the user, a_j are unknown polynomial coefficients (i.e. model parameters) and m is the number of polynomial terms (in addition to the bias term a_0).

Note that the last model structure shown in Equation (1) (i.e. case 3) requires the assumption of an invertible g function because of the subsequent parameter estimation. The set of exponents defined by the user is discrete and should contain zero value. This way, when the exponent $\mathbf{ES}(j,k)$ becomes equal to zero, the value of the kth input variable $\mathbf{X}_k$ in the jth polynomial term is set equal to 1 and that variable is deselected from the model structure. From a statistical point of view this means that variable $\mathbf{X}_k$ is not significant enough to be considered in describing the phenomenon analysed.

The LS method used here (Giustolisi & Savic 2006) provides a two-way correspondence between the model structure and its parameter values. In addition to the unconstrained LS search, the user can force the LS to

search for model structures that contain only positive parameter values ($a_j > 0$) (Lawson & Hanson 1974). This was done since in the modelling of large systems there is a high probability that negative constant value(s) ($a_j < 0$) is/are selected to balance the particular realization of errors related to the finite training data set (Giustolisi *et al.* 2007).

Finally, note that EPR employs a multi-objective search strategy to determine all models that correspond to the optimal trade-off between model parsimony and fitness (Giustolisi *et al.* 2006*b*). Therefore, a single EPR run returns a number of mathematical models (i.e. formulae), each representing a point on the Pareto optimal (accuracy vs. parsimony trade-off) curve of possible models (Pareto 1896).

A model fit to the observed data is evaluated using the Coefficient of Determination (CoD) as follows:

$$\mathrm{CoD} = 1 - \frac{\sum_n (\hat{y} - y_{\exp})^2}{\sum_n (y_{\exp} - \mathrm{avg}(y_{\exp}))^2} = 1 - \frac{n}{\sum_n (y_{\exp} - \mathrm{avg}(y_{\exp}))^2} \cdot \mathrm{SSE} \quad (2)$$

where n is the number of samples, $\hat{y}$ is the value predicted by the model and $\mathrm{avg}(y_{\exp})$ is the average value of the corresponding observations (evaluated on the n samples). Note from Equation (2) that the values of CoD and SSE (sum of squared errors) are strictly correlated, belonging to the same membership of cost functions (Ljung 1999).

Model parsimony is estimated by looking at both the number of polynomial terms and/or the number of input (i.e. explanatory) variables present in the selected model (Giustolisi & Savic 2006). The latest version of the EPR software and methodology allows for the selection of one or both parsimony criteria by performing a two- or three-objective optimization while searching for models (Giustolisi *et al.* 2006*b*).

CASE STUDY

The data in this case study were available at the pipe level for the period 1986–1999 and contain both asset information and recorded bursts. The database used here refers to one of the 48 water quality zones (WQZ) within a UK water distribution system. For each individual pipe, the database contains information on pipe diameter, material, year laid, length, number of properties supplied and the total number of bursts recorded during the 14-year monitoring period. Basic statistics of this data are shown in Table 1. Unfortunately, neither of the criteria adopted for designing this water quality zone nor the network map were available for this study. Furthermore, only the total number of bursts is known (i.e. the timing of each burst is unknown). Lack of the above information prevents verification of the potential existence of spatial and temporal clusters in the burst data.

Table 1 shows that, as in the majority of water distribution systems, the number of bursts recorded during the monitoring period corresponds to less than 10% of the total number of pipes. Furthermore, several pipes failed more than once over the same time period.

It could be argued that, when only failed pipes are considered for developing a statistical model, pertinent results should be referred to as "burst models" since they aim at discovering the causes of failure based on collected information. On the other hand, a "performance indicator" (PI), as it is meant herein, should represent the propensity to fail for all pipes in the network. Such a PI could eventually be used for developing a structural deterioration model to assess individual pipe *criticality* to be considered for decision-making. Therefore, both pipes with and without recorded bursts (Giustolisi & Savic 2004) have been considered here.

As mentioned in the introduction, previous pipe failure models in the literature associated the same pipe burst rate with pipes with similar attributes (e.g. material, size, age, etc.). Following that, and based on the preliminary analyses, the pipes considered here have been classified using pipe diameter and age.

Table 1 | WQZ available pipe features

Features	Values
Year the pipe was laid	From 1910 to 1999
Diameter	From 32 mm to 250 mm
Length	Total 172 984 m
Supplied properties	Total 19 494
Number of pipes	3669
Number of bursts	354

Because the statistical approach is economically viable for modelling failure in small pipes, only pipes with a nominal diameter of up to 250 mm have been selected for the analysis. Chosen pipes have been grouped into 10 diameter classes, with similar classification used to fill-in some existing data gaps. In fact, the completeness of individual records was variable, with numerous missing entries for the year the pipes were laid. In order to fill in these gaps, the correlation often assumed between pipe material and burial year (Pelletier *et al.* 2003) was employed. Thus, within each diameter class, the mean burial year of pipes made of the same material has been used to complete missing data. Once the data reconstruction was completed, pipes were further grouped into one-year-wide age classes.

Only four fields describing pipe features have been considered for modelling. These are *age* (*Ap*), *diameter* (*Dp*), *length* (*Lp*) and number of *properties* (*Pp*) supplied, all available at the pipe level. For each diameter–age class, the total number of recorded burst events (*Brt*), the sum of pipe lengths (*L*), the sum of properties supplied (*P*) and the total number of pipes in the class (*N*) have been computed. Furthermore, to define a significant value of age and diameter for each *class*, the length weighted mean of relevant variables was computed as shown in Equation (3). The values computed are the equivalent age (*A*) and the equivalent diameter (*D*):

$$A_{class} = \frac{\sum_{class}(Lp \cdot Ap)}{L_{class}}; \quad D_{class} = \frac{\sum_{class}(Lp \cdot Dp)}{L_{class}} \tag{3}$$

Note that subscript *class* emphasizes that summation refers to all pipes belonging to the same class. The aforementioned grouping results in a schematization of the network into fictitious pipes whose features are summarized in Figure 1.

The example reported in Table 2 shows in more details the entire grouping procedure starting from a sample dataset. All the information shown in Table 2 is at the pipe level. The six fields report the attributes collected for each pipe: (1) *Pipe ID* – pipe identifier; (2) *Brp* – number of pipe bursts recorded during the monitoring period; (3) *Ap* – pipe age (yr); (4) *Pp* – number of properties supplied; (5) *Lp* – pipe length (m) and (6) *Dp* – pipe nominal diameter (mm). Table 3 reports the same data after the classification using the age and diameter as grouping criteria; diameter classes refer to 63, 75–90 and 100 mm.

In summary, the model under consideration is geared to identifying the functional relationships between five possible model inputs (*A*, *D*, *L*, *N* and *P*) and one model output (*BR*).

DISCUSSION OF DATA AGGREGATION

Since data aggregation plays a significant role in the development of pipe failure models, the following discussion examines some of the important issues, such as the choice of pipe age as grouping criterion and the choice of the equivalent attributes for a pipe class.

A classification based on age allows for indirect consideration of time-varying solicitations on pipes. Although it is impossible to explicitly state the relationship between all solicitation factors and pipe age, engineering knowledge of breakage mechanisms suggests that stress effects (like those that are chemical or mechanical due to soil conditions, traffic loads, variations in service pressure over time, etc.) increase as the duration of solicitation increases. In the case study reported herein, the choice of one year for age classification is due to the following reasons: (i) it averages the influence of time-dependent factors over a year and (ii) it allows for

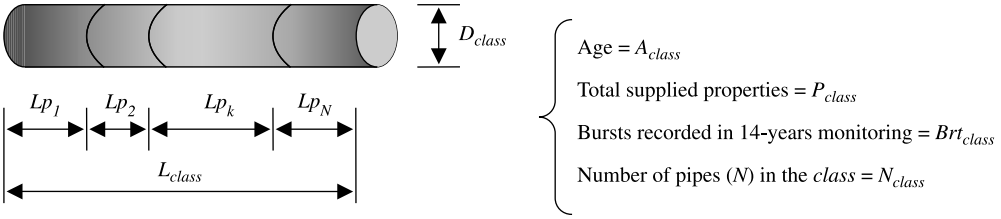

Figure 1 | Features of fictitious pipes representing diameter-age *class*.

Table 2 | Example data

Pipe ID	*Brp*	*Ap*(yr)	*Pp*	*Lp* (m)	*Dp* (mm)
ID 1	1	30	0	10	63
ID 2	3	30	2	55	63
ID 3	0	30	3	35	63
ID 4	0	40	4	10	75
ID 5	5	40	1	55	75
ID 6	2	40	1	5	90
ID 7	2	25	0	10	100
ID 8	3	25	1	35	100
ID 9	4	25	1	15	100

detailed analysis of the problem based on data updated annually by water utilities.

Once aggregation criteria have been selected (e.g. pipe diameter and age), relevant aggregate variables (i.e. equivalent attributes) can be computed as a sum, mean or length-weighted mean (e.g. as for A and D) over each class. Class variables computed as a sum (e.g. L, P and N) are all implicit functions of the classification scheme adopted and their values change if a different aggregation criterion is selected. In particular, the overall class length L has a statistical meaning since it encompasses all other time-related factors that are either unrecorded or unavailable for the same class. For example, the longer the pipe class, the more variable the traffic loads, operational stresses (i.e. pressure/discharge variations) and bedding conditions. Although it is impossible to formulate a mathematical expression of such a relationship without additional information, it is known from the literature that pipe length directly affects the probability of breaks (Jacobs & Karney 1994).

The choice among rationales for computing equivalent attributes reported above (i.e. sum, mean, length-weighted mean) should be consistent with the main schematization of classes as "fictitious" pipes (Figure 1) and with the remaining class variables. For instance, if the length of each class has been computed as a sum over all pipes, the traffic load of the same class, if available, should be computed by summing traffic loads for all pipes as well. Analogously, if information about the pressure regime is available at pipe level, the class pressure should be represented by the length-weighted mean rather than by the arithmetic average of pipe pressures, since the pressure regime affects the entire extent of a pipe. A consistent definition of aggregate variables is advisable in order to have EPR models with a physical meaning.

Table 3 | Data grouped by age (A) and diameter (D)

Class	*Brt*	*A* (yr)	*P*	*L* (m)	*D* (mm)	*N*
1	4	30	5	100	63	3
2	7	40	6	70	76	3
3	9	25	2	60	100	3

EPR SETTINGS

To discover a symbolic relationship between the pipe bursts and grouped pipe attributes, the Case 2 model structure shown in Equation (1) has been used here with function f selected as natural logarithm (Giustolisi & Savic 2006):

$$\mathbf{Y} = \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \cdot \ln\left((\mathbf{X}_1)^{\mathbf{ES}(j,k+1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,2k)}\right) \quad (4)$$

where the input variables are: D, A, P, L and N. The following candidate exponents were considered: $[-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2]$. The model size m (i.e. the number of polynomial terms) was limited to three terms and the bias term was assumed equal to zero. Finally, the LS parameter estimation was constrained to search for positive polynomial coefficient values only ($a_j > 0$).

The absolute values of candidate exponents were chosen to potentially describe linear, square or even half-power functions, while their positive and negative signs allow for the representation of direct and inverse relationships between inputs and the output.

The natural logarithm was selected for a possible functional transformation f to test if the relationships between input variables and the output could benefit from using two different scales in a single equation (Savic *et al.* 2006). The results have subsequently shown that, in the most meaningful models (from the engineering point of view), the natural logarithm in Equation (4) had not been selected as all the exponents were found to be zero.

MODEL SELECTION

Once applied, the (single) EPR run returned a set of burst prediction models as a Pareto set, trading off model parsimony with a fit to the observed data. Table 4 lists the one- and two-term models obtained, while Figures 2 and 3 show these models as points in the objective space. More precisely, Figures 2 and 3 represent the projections of the overall Pareto front on the corresponding objective planes. The plane in Figure 2 is defined by the objectives representing the number of polynomial terms (i.e. number of a_j) and model fit. The plane in Figure 3 is defined by the objectives representing the number of model input variables Xi and model fit to the observed data. Note that in both figures model fit is calculated as 1-CoD to draw a Pareto front corresponding to a minimization problem.

All models shown in Table 4 clearly demonstrate that burst occurrence depends only on the following three (out of five) candidate input variables: the equivalent pipe class age A, the equivalent pipe class diameter D and the total pipe class length L. The inverse dependence between diameter and burst occurrence is confirmed by all models as well as is the direct dependence on the class length and equivalent age. Note that only the first two models do not have all three significant input variables. In particular, the first model shows that the number of connections P describes about 55% of burst variation among classes, while the second one reports a significant improvement in terms of CoD when L and D are considered only. It is evident that the selection of variable A further improves model performance. Introducing other variables or even a second polynomial term does not improve significantly the model fit. Bearing in mind the above discussion and incorporating engineering insight into the problem, the following model is selected:

$$BR = 0.084904 \cdot \frac{A \cdot L}{D^{1.5}} \tag{5}$$

Units are yr, m and mm for variables A, L and D, respectively. The above model fits the observed data with CoD = 0.822 and is represented in Figure 4. Note that the selected model is located near the inflection point on Pareto fronts shown in Figures 2 and 3 (indicated by an arrow), implying that this model ensures substantial improvement in the model fit with only little increase in model complexity.

The chosen model highlights that, for the analysed water distribution system, pipe age and diameter are important, but so too is pipe length. This confirms previous findings in most of the literature on the subject, as discussed in the introduction section. In particular, the linear relationship between the number of pipe bursts and pipe age should be ascribed to the fact that the system is

Table 4 | Formulae retuned by EPR

EPR formula	CoD	No. of a_j	No. of X_h
$BR = 2.7832 \times 10^{-5}.P^2$	0.550	1	1
$BR = 0.045514(L^{1.5}/D^{1.5})$	0.704	1	2
$BR = 0.084904(AL/D^{1.5})$	0.822	1	3
$BR = 1.1895 \times 10^{-5} \cdot P^2 + 0.012065(A^2L/D^2)$	0.842	2	4
$BR = 0.019882(AL/D^{1.5})\ln(L^{0.5})$	0.822	1	4
$BR = 0.00013397(P^2/D^{0.5}) + 0.011772(A^2L/D^2)$	0.842	2	5
$BR = 0.013684(AL/D^{1.5})\ln(A^{0.5}L^{0.5})$	0.838	1	5
$BR = 1.0049 \times 10^{-5}P^2 + 0.0083887(A^{1.5}L/D^2)\ln(L^2/P)$	0.857	2	6
$BR = 0.02184(AL/D^{1.5})\ln(A^{0.5}L^{0.5}/D^{0.5})$	0.827	1	6
$BR = 5.3118 \times 10^{-6}P^2\ln(A^{0.5}) + 0.0083637 \cdot (A^{1.5}L/D^2)\ln(L^2/P)$	0.854	2	7
$BR = 0.0037636(P^{1.5}/D^{0.5}) + 0.0010914(A^2L/D^2)\ln(L^{1.5}N/P)$	0.854	2	8
$BR = 0.002062(P^{1.5}/D^{0.5})\ln(A^{0.5}) + 0.0010745(A^2L/D^2)\ln(L^{1.5}N/P)$	0.862	2	9
$BR = 0.00060897(A^{0.5}P^{1.5}/D^{0.5}) + 0.0013296(A^2L/D^2)\ln(L^{1.5}N/PD^{0.5})$	0.861	2	10
$BR = 0.00063284(A^{0.5}P^{1.5}/D^{0.5}) + 0.0013552(A^2L/D^2)\ln(A^{0.5}L^{1.5}N/PD)$	0.859	2	11

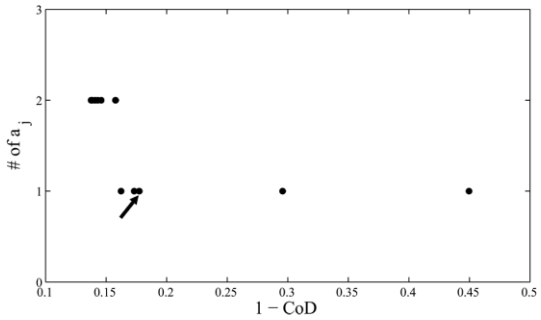

Figure 2 | Projection of Pareto front of models on the plane: Fitness–Number of terms.

(on average) experiencing a wear-out phase on the so-called bathtub curve (Andreou *et al.* 1987*a*, *b*; Kleiner & Rajani 2001; Watson 2005). As reported previously (Kettler & Goulter 1985; Zhao 1998), pipe diameter plays an important role too, indicating that smaller diameter pipes are more prone to failing under excessive external stresses than larger ones. This behaviour could be due to numerous reasons including pipe manufacturing issues and/or typically low quality of workmanship involved when installing small diameter pipes. Equation (5) confirms that the longer is the class (i.e. individual pipe) the higher is the number of bursts (Jacobs & Karney 1994). It is noteworthy that the linear relation between pipe length and number of bursts is a result of EPR analysis rather than a hypothesis.

DERIVING PERFORMANCE INDICATORS USING EPR

The IWA Manual of Best Practice on *Performance Indicators for Water Supply Services* (Alegre *et al.* 2000) describes a performance indicator as a *quantitative measure of a particular aspect of the undertaking's performance or standard of service* assisting *in the monitoring and evaluation of the efficiency and effectiveness of the undertakings thus simplifying an otherwise complex evaluation*. Among the rationales suggested for establishing whether a certain quantity can be considered as a Performance Indicator are the following criteria: (1) to be easy to understand even by non-specialists; (2) to be applicable to undertakings with different characteristics and stages of development and (3) to be as few as possible, avoiding the inclusion of non-essential aspects.

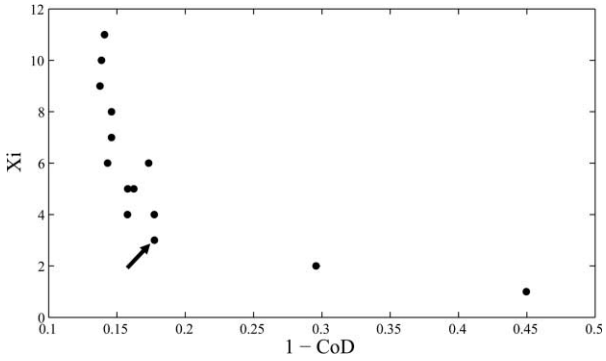

Figure 3 | Projection of Pareto front of models on the plane: Fitness–Number of selected inputs.

The number of pipe failures as a performance indicator has been included in the above manual as the number of failures per 100 km of pipeline per year. The case study reported here shows how the EPR modelling technique provides a tool that can formulate such an indicator as a function of the simplest asset features of the system (i.e. length, diameter and age). The entire methodology from the raw data analysis to the model shown in Equation (5) satisfies each of the above criteria, thus indicating that EPR can be used for analysing performance indicators for water systems. The main reasons for this are as follows:

1. The data manipulation is simple and the search for the pipe burst model is basically described as the best combination of input variables' exponents.
2. The whole methodology could be applied to undertakings with different characteristics (e.g. age, pipe material composition, pressure regime, etc.). In fact, the user could select the same temporal bounds and the same classification criteria for data belonging to different systems. The resulting models, expressed in a compact form like that in Equation (5), could be used for both assessing the number of pipe failures and for finding the most influential variables among those selected as inputs. Furthermore, the model obtained for a given system could be valid for similar systems, apart from the scaling factor (i.e. first constant of the formula). This is due to the fact that EPR formulae are symbolic and that the search method explicitly avoids over-fitting the data, thus allowing the description of the physical phenomenon (Berardi & Kapelan 2006). Moreover, the EPR methodology can be used on datasets corresponding to either the entire system or some subsystems (Berardi *et al.* 2005).

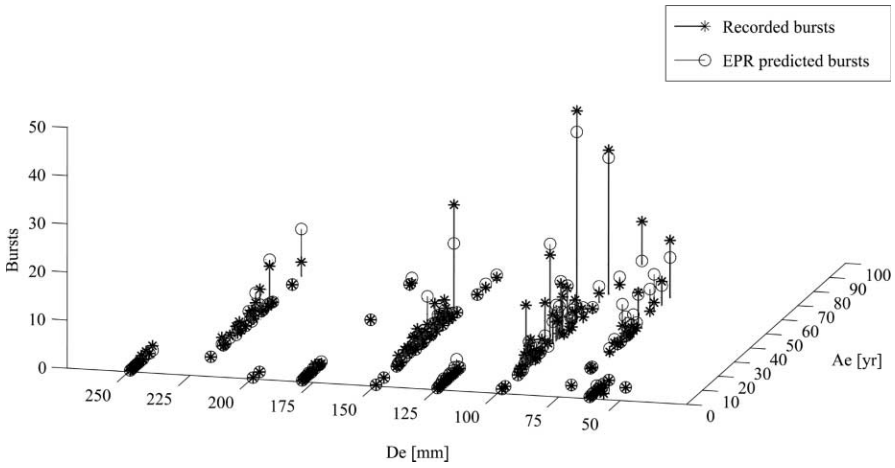

Figure 4 | Fitting for burst distribution model.

3. Both the pre-processing and the modelling phases are aimed at avoiding inclusion of non-essential aspects of the phenomenon modelled. In particular, during the pre-processing phase (i.e. data classification and input selection) the user is allowed to choose between available input variables. During the modelling phase, EPR employs a multi-objective search that is able to deselect (i.e. by assigning an exponent equal to zero) those input variables which are not required to describe the phenomenon analyzed. Moreover, each of the expressions returned by the EPR is evaluated in terms of its CoD, thus making the user aware of the reliability of information obtained.

PIPE DETERIORATION MODELS BASED DECISION SUPPORT

A DSS for water distribution pipe rehabilitation/replacement should include the assessment of pipe *criticality* in terms of its failure risk, where the risk is defined as the product of pipe failure likelihood/frequency (e.g. number of predicted bursts per unit time) and the expected damage due to failure (i.e. pipe burst). The damage term should take into account both direct (i.e. repair) and indirect (environmental, social and third party) costs. In particular, the introduction of some "damage multiplier" (Walski & Pelliccia 1982) or "cost factors" (Dandy & Engelhardt 2006) could explicitly account for additional information such as network connectivity or land use.

The number of bursts should normally be assessed for each single pipe over a given time horizon. As the EPR models reported here are aggregated, they cannot be used directly for assessing burst rate at the individual pipe level. Previous research on EPR modelling (Berardi *et al.* 2005; Giustolisi *et al.* 2006*a*) reports pipe length as the main criterion for assessing individual pipe burst rate. In more general terms, all aggregated variables found in an EPR model should be considered. Given an observation period T, the burst rate λ_i^{EPR} for pipe i belonging to a given *class* can be calculated as follows:

$$\lambda_i^{EPR} = \frac{BR_{class}}{T} \cdot \frac{\sum_s \frac{Vp_{i,s}}{V_{class,s}}}{\sum_{class} \sum_s \frac{Vp_{i,s}}{V_{class,s}}} \tag{6}$$

where s is the index of the sth aggregate variable V selected in the EPR model, subscripts i and *class* emphasize that such a variable refers to either the ith pipe or the entire *class* the pipe belongs to. In the case of model (5), the aggregate variables considered are L, P and N, all computed by summation, but only class length L has been selected in the EPR model. Thus Equation (6) can be written as follows:

$$\lambda_i^{EPR} = \frac{BR_{class}}{T} \cdot \frac{\frac{Lp_i}{L_{class}}}{\sum_{class} \frac{Lp_i}{L_{class}}} = \frac{BR_{class}}{T} \cdot \frac{Lp_i}{\sum_{class} Lp_i} = \frac{BR_{class}}{T} \cdot \frac{Lp_i}{L_{class}} \tag{7}$$

where Lp_i is the length of pipe i.

It is worth mentioning that λ_i^{EPR} represents the burst rate (bursts per year) of pipe i due to its membership to the *class* and does not take into account its individual burst history. Nonetheless, information about individual pipe

burst history is of great relevance for establishing an individual pipe's propensity to fail, especially for supporting rehabilitation or replacement decisions.

Information on individual pipe history can be quantified as the ratio between the number of burst events experienced by pipe i and the relevant observation period T:

$$\lambda_i^R = \frac{Brp_i}{T} \tag{8}$$

Without any additional information, such a "recorded" burst rate λ_i^R can be assumed constant over the observation period T and equal to 0 for all those pipes which did not experience any burst during the same interval.

In order to account for individual burst history a general structural deterioration model based on the EPR aggregate model is developed here.

It can be argued that burst rate λ_i^{EPR} ((6) and (7)) depends on time since BR_{class} itself is a function of age variable A. Thus, after t years from the end of the observation period, λ_i^{EPR} should be calculated as follows:

$$\lambda_i^{EPR}(t) = \frac{\left\{BR_{class} = a_1 \cdot \left[D_{class}^{\delta} \cdot L_{class}^{\gamma} \cdot P_{class}^{\rho} \cdot N_{class}^{\mu} \cdot (A_{0,class} + t)^{\alpha}\right]\right\}}{T} \cdot \frac{Lp_i}{L_{class}} \tag{9}$$

where δ, γ, ρ, μ and α represent the exponents selected in the EPR model for variables D, L, P, N and A, respectively. The exponent α in Equation (9) can assume both positive and negative values depending on the model structure returned by EPR. The value of α can be linked to a particular deterioration phase of the system, as shown in Figure 5. In the case of model (5), Equation (9) becomes as follows:

$$\lambda_i^{EPR}(t) = \frac{1}{T} \cdot a_1 \frac{L_{class}(A_{0,class} + t)}{D_{class}^{1.5}} \cdot \frac{Lp_i}{L_{class}} = \frac{a_1}{T} \cdot \frac{Lp_i(A_{0,class} + t)}{D_{class}^{1.5}} \tag{10}$$

where a_1 denotes the corresponding EPR model coefficient (e.g. $a_1 = 0.084904$ in the case study reported here) and $A_{0,class}$ is the equivalent age of the class when $t = 0$ (i.e. at the end of the monitoring period).

It is worth noting that, in this case, the dependence on time t can be expressed explicitly for variable A only (i.e. $A = A_0 + t$), while the other variables are assumed to be constant over time. However, if other time-dependent variables (e.g. traffic load, soil moisture and so on) were available and selected by EPR, they could be included into the analyses by incorporating relevant functional relations into Equations (9) and (10). This way, such a formulation allows for describing pipe ageing by including other variables, even different from age A.

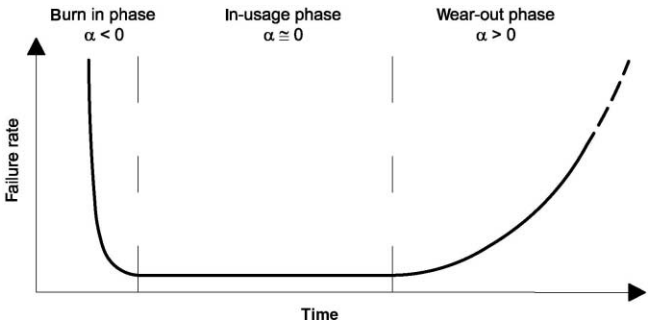

Figure 5 | Bathtub curve.

The model in Equation (9) and (10) could be potentially used for all pipes in the network. Nevertheless, the behaviour of a pipe which experienced one or more burst events during the observation period T can be described better by its own observed burst rate λ_i^R (8). The observed failure rate λ_i^R for pipe i can be written as follows:

$$\lambda_i^R(t) = \frac{Brp_i}{T} = \frac{a_i \cdot \left[D_{class}^{\delta} \cdot L_{class}^{\gamma} \cdot P_{class}^{\rho} \cdot N_{class}^{\mu} \cdot (A_{0,class} + t)^{\alpha}\right]}{T} \cdot \frac{Lp_i}{L_{class}} \tag{11}$$

where coefficient a_i is computed at planning time $t = 0$ as in Equation (12):

$$a_i = \frac{Brp_i}{D_{class}^{\delta} \cdot L_{class}^{\gamma} \cdot P_{class}^{\rho} \cdot N_{class}^{\mu} \cdot A_{0,class}^{\alpha}} \cdot \frac{L_{class}}{Lp_i} \tag{12}$$

Obviously, different coefficients are computed for all failed pipes even if they belong to the same *class*. Both coefficients a_1 and a_i are expressed using the same units, which depend on the main EPR model structure. In the case reported in Equation (5), they are expressed in terms of $mm^{-1.5}\ m^{-1}\ yr^{-1}$.

In summary, given the pipe i belonging to a particular *class*, its failure rate can be calculated as follows:

$$\lambda_i(t) = \begin{cases} \frac{a_1 \cdot \left[D_{class}^{\delta} \cdot L_{class}^{\gamma} \cdot P_{class}^{\rho} \cdot N_{class}^{\mu} \cdot (A_{0,class} + t)^{\alpha}\right]}{T} \cdot \frac{Lp_i}{L_{class}} & \text{if } Brp_i = 0 \\ \frac{a_i \cdot \left[D_{class}^{\delta} \cdot L_{class}^{\gamma} \cdot P_{class}^{\rho} \cdot N_{class}^{\mu} \cdot (A_{0,class} + t)^{\alpha}\right]}{T} \cdot \frac{Lp_i}{L_{class}} & \text{if } Brp_i > 0 \end{cases} \tag{13}$$

This means that the predicted number of bursts for individual pipes is obtained by combining model structure and coefficient estimation. The information about the individual pipe burst history is used to improve the accuracy of the predicted burst rate while employing the same deterioration model structure as that returned by EPR. In the case of pipes without a documented burst history the model coefficient is assumed to be the same as that returned by EPR (i.e. a_1), whereas, for pipes which experienced burst events during the monitoring period, model coefficient a_i is computed from their own burst history.

In the case of the water distribution network reported here, Equation (13) becomes

$$\lambda_i(t) = \begin{cases} \frac{a_1}{T} \cdot \frac{Lp_i(A_{0,class}+t)}{D_{class}^{1.5}} & \text{if} \quad Brp_i = 0 \\ \frac{a_i}{T} \cdot \frac{Lp_i(A_{0,class}+t)}{D_{class}^{1.5}} & \text{if} \quad Brp_i > 0 \end{cases} \tag{14}$$

where coefficient a_i for failed pipes is computed as follows:

$$a_i = Brp_i \cdot \frac{D_{class}{}^{1.5}}{A_{0,class}} \cdot \frac{1}{Lp_i} \tag{15}$$

Formulation (13) of the failure rate is used to predict the individual number of bursts (BR_i) in a given planning horizon h as in Equation (16):

$$BR_i = \int_0^h \lambda_i(t) \cdot dt \tag{16}$$

Once the number of bursts predicted, BR_i, has been computed for all pipes, it can be used with damage d_i caused by a single burst event to calculate the risk of burst. A decision support methodology can then be employed to select those pipes with the highest risk value R_i:

$$R_i = d_i \cdot BR_i \tag{17}$$

The methodology presented above leads to the over-estimation of the number of bursts for those pipes without a failure history, as the best prediction for them is zero burst at $t = 0$. In order to overcome this drawback, additional information on individual pipe burst history could be introduced into the decision support system. This way, the decision support problem is defined as a multi-objective optimization problem (Giustolisi *et al.* 2006*a*) by using the following three objectives: minimize the cost of interventions (e.g. pipe replacement or refurbishment) whilst maximizing (for selected pipes) the risk functions based on both future burst prediction and individual failure history.

Finally, it is worth remarking that the calculation of a burst rate as in Equations (9) and (13) deals mainly with small pipes that the EPR model has been developed for. In the case of large size pipes consequences of failure are large enough to justify careful inspection and the development of physically based models (Kleiner & Rajani 2001). Having said this, as in the case of small size pipes, the risk of large size pipes (e.g. trunk mains) failures can also be estimated (by quantifying the corresponding failure probabilities and the associated damages) and used in the DSS.

CONCLUSIONS

An application of a new data mining technique for failure prediction in water distribution systems is described in this paper. The technique, called Evolutionary Polynomial Regression (EPR), produces symbolic expressions that are essentially explicit mathematical models for pipe burst predictions originating in data-driven analysis. Unlike other data mining techniques, EPR produced simple and understandable relationships/models that provide a high level of statistical correlation among the variables. The models obtained can also be validated by means of physical knowledge.

The approach is tested and verified on a real-life UK water distribution system. The case study demonstrates the entire process, from data aggregation to EPR model selection. The resulting EPR aggregate model is effective in terms of regression performance (CoD) and is consistent with the physical/engineering understanding of the problem. Pipe age, diameter and length have been selected as the most important variables in describing pipe burst occurrence and direct/inverse relations confirm previous findings on the subject. Because of these characteristics the relationships obtained by using EPR have been regarded as performance indicators of the network and the possible employment of the EPR technique in developing other performance indicators for water systems has been discussed.

Finally, an individual pipe structural deterioration model has been derived from the EPR aggregate model

and the methodology on how to use such a model to support asset management decision-making is presented.

REFERENCES

Alegre, H., Hirnir, W., Baptista, J. M. & Parena, R. 2000 *Performance Indicators for Water Supply Services. Manual of Best Practice*. IWA Publishing, London.

Andreou, S. A., Marks, D. H. & Clark, R. M. 1987*a* A new methodology for modelling break failure patterns in deteriorating water distribution systems: theory. *Adv. Wat. Res.* **10**, 2–10.

Andreou, S. A., Marks, D. H. & Clark, R. M. 1987*b* A new methodology for modelling break failure patterns in deteriorating water distribution systems: applications. *Adv. Wat. Res.* **10**, 11–20.

Babovic, V., Drécourt, J., Keijzer, M. & Hansen, P. F. 2002 A data mining approach to modelling of water supply assets. *Urban Wat.* **4**, 401–414.

Berardi, L., Savic, D. A. & Giustolisi, O. 2005 Investigation of burst-prediction formulas for water distribution systems by evolutionary computing. In *Proc. of the 8th International Conference on Computing and Control for the Water Industry*, (ed. D. A. Savic, G. A. Walters, R. King & S.-T. Khu), University of Exeter, UK, CCWI2005, Vol. 2, pp. 275–280.

Berardi, L. & Kapelan, Z. 2007 Multi-case EPR strategy for the development of sewer failure performance indicators. *Proc. World Environmental and Water Resources Congress*, (K. C. Kabbes, Ed.), ASCE, Reston, VA. doi: 10.1061/40927(243)162.

Bessler, F. T., Savic, D. A. & Walters, G. A. 2002 Pipe burst risk analysis with data mining. In *Proc. of the 5th International Conference on Hydroinformatics, Hydroinformatics, 1–5 July, Cardiff*, (ed. I. D. Cluckie, D. Han, J. P. Davis & S. Heslop), IWA Publishing, London, Vol. 1, pp. 783–788.

Clark, R. M., Stafford, C. L. & Goorich, J. A. 1982 Water distribution systems: a spatial and cost evaluation. *J. Wat. Res. Plann. Manage. Div.* **108** (3), 243–256.

Constantine, A. G., Darroch, J. N. & Miller, R. 1996 Predicting underground pipe failure. *Water (J. Australian Wat. Assoc.)* **23** (2), 9–10.

Dandy, G. C. & Engelhardt, M. O. 2006 Multi-objective trade-offs between cost and reliability in the replacement of water mains. *J. Wat. Res. Plann. Manage.* **132** (2), 79–88.

Draper, N. R. & Smith, H. 1998 *Applied Regression Analysis*, 3rd edn. John Wiley & Sons, New York.

EPR web site. Available at: http://www.hydroinformatics.it.

Farley, M. & Trow, S. 2003 *Losses in Water Distribution Networks – A Practitioner's Guide to Assessment, Monitoring and Control*. IWA Publishing, London.

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. 1996 From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining* (ed. Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth & Ramasamy Uthurusamy), AAAI Press and MIT Press, California, pp. 1–34.

Giustolisi, O. 2004 Using genetic programming to determine Chézy resistance coefficient in corrugated channels. *J. Hydroinf.* **6** (3), 157–173.

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

Giustolisi, O. & Savic, D. A. 2004 Decision support for water distribution system rehabilitation using evolutionary computing. In *Proc. of the Seminar on Decision Support in the Water Industry under Conditions of Uncertainty*, (ed. G. Walters, D. Savic, S. -T. Khoo & R. King), University of Exeter, UK, pp. 76–83.

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

Giustolisi, O., Laucelli, D. & Savic, D. A. 2006*a* Development of rehabilitation plans for water mains replacement considering risk and cost-benefit assessment. *Civ. Engng. Environ. Syst. J.* **23** (3), 175–190.

Giustolisi, O., Savic, D. A. & Kapelan, Z. 2006*b* Multi-objective evolutionary polynomial regression. In *Proc. of the 7th International Conference on Hydroinformatics*, (ed. P. Gourbesville, J. Cunge, V. Guinot & S. Y. Liong), Research Publishing, India, HIC 2006, 4–8 September, Nice, France, Vol. 1, pp. 725–732.

Goulter, I. C. & Kazemi, A. 1988 Spatial and temporal groupings of water main pipe breakage in Winnipeg. *Can. J. Civ. Eng.* **15** (1), 91–97.

Herz, R. K. 1996 Ageing processes and rehabilitation needs of drinking water distribution networks. *JWSRT-Aqua* **45** (5), 221–231.

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

Kettler, A. J. & Goulter, I. C. 1985 An analysis of pipe breakage in urban water distribution networks. *Can. J. Civ. Eng.* **12** (2), 286–293.

Kleiner, Y. & Rajani, B. B. 1999 Using limited data to assess future needs. *J. AWWA* **91** (7), 47–62.

Kleiner, Y. & Rajani, B. B. 2001 Comprehensive review of structural deterioration of water mains: statistical models. *Urban Wat.* **3** (3), 121–150.

Kleiner, Y. & Rajani, B. B. 2002 Forecasting variations and trends in water-main breaks. *J. Infrastruct. Syst.* **8** (4), 122–131.

Koza, J. R. 1992 *Genetic Programming: On the Programming of Computers by Natural Selection*. MIT Press, Cambridge, MA.

Lawson, C. L. & Hanson, R. J. 1974 *Solving Least Squares Problems*, Englewood Cliffs, NJ, Prentice-Hall. ch. 23, pp. 161.

Le Gat, Y & Eisenbeis, P. 2000 Using maintenance records to forecast failures in water network. *Urban Wat.* **2** (3), 173–181.

LeGauffre, P., Baur, R., Laffrenchine, K. & Miramond, M. 2002 Multi-criteria decision aid for rehabilitation of water supply networks. In *Proc. 3rd International Conference on Decision Making in Urban and Civil Engineering*, DMUCE, London, UK, pp. 655–660.

Lei, J. & Saegrov, S. 1998 Statistical approach for describing failures and lifetime of water mains. *Wat. Sci. Technol.* **38** (6), 209–217.

Ljung, L. 1999 *System Identification: Theory for the User*, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ.

Matos, R., Cardoso, A., Ashley, R. M., Molinari, A., Schulz, A. & Duarte, P. 2003 *Performance Indicators for Wastewater Services. IWA Manual of Best Practice*. IWA Publishing, London.

Mavin, K., 1996 Predicting the failure performance of individual water mains. *Urban Water Research Association of Australia, Research Report No.* 114. Melbourne, Australia.

McDonald, S. E. & Zhao, J. Q. 2001 Condition assessment and rehabilitation of large sewers. In *Proc. International Conference on Underground Infrastructure Research*, pp. 361–369. University of Waterloo, Waterloo, Ontario, 10–13 June.

Pareto, V. 1896 *Cours d'economie politique*, Rouge & Cic. Lausanne, Switzerland Vols. 1 and 2.

Pelletier, G., Mailhot, A. & Villeneuve, J. P. 2003 Modeling water pipe breaks–three case studies. *J. Wat. Res. Plann. Mngmnt.* **129** (2), 115–123.

Savic, D. A., Djordjevic, S., Dorini, G., Shepherd, W., Cashman, A. & Saul, A. 2005 COST-S: A new methodology and tools for sewerage asset management based on whole life costs. *Wat. Asset Manage. Int.* **1** (4), 20–24.

Savic, D. A., Giustolisi, O., Berardi, L., Shepherd, W., Djordjevic, S. & Saul, A. 2006 Modelling sewers failure using evolutionary computing. *Proc. ICE, Wat. Manage.* **159** (2), 111–118.

Shamir, U. & Howard, C. D. D. 1979 An analytic approach to scheduling pipe replacement. *J. AWWA* **117** (5), 248–258.

Shepherd, W., Cashman, A., Djordjevic, S., Dorini, G., Saul, A., Savic, D. A., Lewis, L., Walters, G. & Ashley R. 2004 Whole life costing of sewer systems. *WaPUG Autumn Meeting, Blackpool, 10–12 Nov*. Available at: http://www.wapug.org.uk/past_papers/Autumn_2004/shepherd.pdf.

Skipworth, P., Engelhardt, M., Cashman, A., Savic, D. A., Saul, A. J. & Walters, G. A. 2002 *Whole Life Costing for Water Distribution Network Management*. Thomas Telford, London.

Walski, T. M. & Pelliccia, A. 1982 Economic analysis of water main breakes. *J. AWWA* **74** (3), 140–147.

Watson, T. 2005 A hierarchical Bayesian model and simulation software for water pipe networks. *Civil Engineering*. The University of Auckland, Auckland.

Zhao, J. Q. 1998 Trunk sewers in Canada. In: *APWA International Public Works Congress Seminar Series*, American Public Works Association, Las Vegas, pp. 75–89.

First received 29 June 2006; accepted in revised form 19 August 2007