Asset deterioration analysis using multi-utility data and multi-objective data mining

D. A. Savic, O. Giustolisi and D. Laucelli

ABSTRACT

Physically-based models derive from first principles (e.g. physical laws) and rely on known variables and parameters. Because these have physical meaning, they also explain the underlying relationships of the system and are usually transportable from one system to another as a structural entity. They only require model parameters to be updated. Data-driven or regressive techniques involve data mining for modelling and one of the major drawbacks of this is that the functional form describing relationships between variables and the numerical parameters is not transportable to other physical systems as is the case with their classical physically-based counterparts. Aimed at striking a balance, Evolutionary Polynomial Regression (EPR) offers a way to model multi-utility data of asset deterioration in order to render model structures transportable across physical systems. EPR is a recently developed hybrid regression method providing symbolic expressions for models and works with formulae based on pseudo-polynomial expressions, usually in a multi-objective scenario where the best Pareto optimal models (parsimony versus accuracy) are selected from data in a single case study. This article discusses the improvement of EPR in dealing with multi-utility data (multi-case study) where it has been tried to achieve a general model structure for asset deterioration prediction across different water systems.

Key words | asset deterioration, data mining, evolutionary computing, sewer, water supply networks

D. A. Savic (corresponding author)
Centre for Water Systems, School of Engineering, Computing and Mathematics,
University of Exeter,
Harrison Building,
North Park Road,
Exeter EX4 4QF,
UK
E-mail: *d.savic@exeter.ac.uk*

O. Giustolisi
D. Laucelli
Department of Civil and Environmental Engineering,
Technical University of Bari,
II Engineering Faculty,
Taranto via Turismo 8, 74100,
Italy

INTRODUCTION

A number of different performance indicators (PIs) for the sewer and water systems have been proposed recently (Alegre *et al.* 2000; Matos *et al.* 2003; Tran *et al.* 2007). PIs provide regulatory and policy-making bodies with a common basis for measuring and comparing the performance of different drinking and wastewater utilities and identifying possible corrective measures as part of proactive system management. In the case of sewer systems, for example, the proactive approach is aimed at locating the critical pipe sections that need repair or replacement so that relevant pipe inspections and maintenance/rehabilitation works can be planned (Ariaratnam *et al.* 2001). Equally, in water distribution systems, the prediction of deterioration can be used for the development of strategies for water mains replacement considering risk and cost-benefit assessment (Babovic *et al.* 2002; Giustolisi *et al.* 2006).

The technical literature on sewer PIs reveals two main approaches. The first exploits technical expertise gained from the management of real networks and seeks to define a set of indicators shared by as many utilities as possible. Studies following this approach suggest a list of rationales for establishing whether a certain parameter can be considered as a Performance Indicator (Alegre *et al.* 2000). The second approach aims at developing PIs from hydraulic (Cardoso *et al.* 1999) and asset performance (Berardi *et al.* 2005; Savic *et al.* 2006). These models are

doi: 10.2166/hydro.2009.019

based on the analysis of existing company databases which archive data on sewer assets and preserve the historical records of failure events. The scope of such analyses is discovering patterns in asset data for describing pipe failures (e.g. sewer blockages or collapses and water distribution pipe bursts). Since it is a data-driven approach, a preliminary overview of commonly available databases is required to select the most effective modelling technique to be used.

Real-life asset and failure datasets tend to differ in both quality and quantity and are stored in separate databases without adequate references to link them. Moreover, the shortness of a monitoring period often results in a small number of recorded failure events. This scarcity of historical data only permits assessment of a finite individual failure probability for a small fraction of pipes. In the case of water distribution networks, an intuitive solution for this consists of aggregating pipes into homogeneous groups (Shamir & Howard 1979). This way, the lack of data problem is overcome and unreliable information is averaged over groups. Pipe grouping allows also for the assessment of an individual pipe failure probability by assuming the same behaviour for similar pipes.

The selection of meaningful grouping criteria is strongly conditioned by the quality and type of data available. In general, pipe failure data is either available at the single pipe level or at a group pipe level (e.g. geographical area). The former enables the definition of the pipe grouping criteria based on various potential explanatory variables (Savic *et al.* 2006). The latter does not allow any further rationales to be included in data preparation and relevant groups can be based on topological proximity criteria only (Berardi *et al.* 2006).

Once pipe groups have been defined, an effective modelling technique is needed to highlight the most significant explanatory variables for describing the deterioration phenomenon. Berardi *et al.* (2005) and Savic *et al.* (2006) demonstrated, for waste and clean water systems, the effectiveness of using the Evolutionary Polynomial Regression (EPR) technique (Giustolisi & Savic 2006) for discovering patterns between failure numbers and potential explanatory variables (e.g. pipe age, size, gradient, etc.). General validity of individual system failure models returned by EPR was checked only by means of a cross correlation analysis; that is, by applying the failure model developed for one system to predict failures in another (Berardi *et al.* 2006). The results obtained in the case of blockage prediction models led to the conclusion that EPR could potentially be used for developing failure models of general validity.

Still, despite the promise of wider applications, the EPR methodology presents some drawbacks when developing generalized asset failure models. These are as follows: (1) failure models developed for different individual systems usually differ in one or more significant explanatory variables or, sometimes, even in model structure (entire polynomial term(s)); (2) it is difficult to establish which of the failure models returned for individual systems should be used as a generalized model (i.e. a performance indicator model). These weaknesses are expected to be intensified as the number of systems analyzed increases.

This paper proposes a novel methodological approach where the EPR technique is used to develop generalized pipe failure prediction models by simultaneously considering pipe failure and attribute data from a number of individual systems. The approach is entitled the Multi-Case Strategy for EPR (MCS-EPR).

BRIEF INTRODUCTION TO EPR

Numerical regression is the most powerful and commonly applied form of regression that provides a solution to the problem of finding the best model to fit the observed data (e.g. fitting a line through a set of points). However, the functional form (linear, exponential, logarithmic, etc.) has to be selected before fitting commences. On the other hand, genetic programming uses simple, but very powerful artificial intelligence tactics for computer learning inspired by natural evolution to find the appropriate mathematical model to fit a set of points. The computer produces and evolves a whole population of functional expressions based on how closely each of them fit the data. The automated induction of mathematical models (descriptions) of data using genetic programming (Koza 1992) is commonly referred to as symbolic regression (Babovic & Keijzer 2000). Evolutionary Polynomial Regression (EPR) is a recently developed hybrid regression method by

Giustolisi & Savic (2006) that integrates the best features of numerical regression (Draper & Smith 1998) with genetic programming (Koza 1992).

The general expression of the EPR formula is given as

$$\hat{\mathbf{Y}} = \sum_{j=1}^{m} F\left(\mathbf{X}, f(\mathbf{X}), a_j\right) + a_0 \tag{1}$$

where $\hat{\mathbf{Y}}$ is the estimated output of the system/process; a_j is a constant value; F is a function constructed by the process; $\mathbf{X}$ is the matrix of input variables; f is a function defined by the user; and m is the length (number of terms) of the polynomially structured expression (bias a_0 excluded, if any) (Giustolisi & Savic 2006).

EPR works with a Genetic Algorithm (GA) (Holland 1975) that is developed *ad hoc* (Giustolisi *et al.* 2004). Moreover, the combination of the GA for finding the best function structures F and LS (Least Squares) for the identification of the constant values a_j offers many advantages. On the one hand, a two-way (biunique) relationship between the model structure and constants is guaranteed by LS; on the other, the GA performs a global exploration of the model space (symbolic expressions) in a single/multi-objective function scenario. The LS strategy for a_j is usually robust with respect to the number of parameters required by EPR and amount of measured data available.

Finally, EPR allows pseudo-polynomial expressions belonging to the class of Equation (1) such as

$$\begin{aligned}
\hat{\mathbf{Y}} &= a_0 + \sum_{j=1}^{m} a_j (\mathbf{X}_1)^{\mathrm{ES}(j,1)} \ldots (\mathbf{X}_k)^{\mathrm{ES}(j,k)} f((\mathbf{X}_1)^{\mathrm{ES}(j,k+1)}) \\
&\quad \times \ldots f((\mathbf{X}_k)^{\mathrm{ES}(j,2k)}) \\
\hat{\mathbf{Y}} &= a_0 + \sum_{j=1}^{m} a_j f((\mathbf{X}_1)^{\mathrm{ES}(j,1)} \ldots (\mathbf{X}_k)^{\mathrm{ES}(j,k)}) \\
\hat{\mathbf{Y}} &= a_0 + \sum_{j=1}^{m} a_j (\mathbf{X}_1)^{\mathrm{ES}(j,1)} \ldots (\mathbf{X}_k)^{\mathrm{ES}(j,k)} f((\mathbf{X}_1)^{\mathrm{ES}(j,k+1)} \\
&\quad \times \ldots (\mathbf{X}_k)^{\mathrm{ES}(j,2k)}) \\
\hat{\mathbf{Y}} &= g\left(a_0 + \sum_{j=1}^{m} a_j (\mathbf{X}_1)^{\mathrm{ES}(j,1)} \ldots (\mathbf{X}_k)^{\mathrm{ES}(j,k)}\right)
\end{aligned} \tag{2}$$

where $\hat{\mathbf{Y}}$ is the vector of model predictions and k is the number of candidate-independent variables or inputs. User-specified functions f reported in Equations (1) and (2) may be natural logarithmic, exponential, tangent hyperbolic, etc. Note that the last structure in Equations (2) requires the assumption of an invertible function g, because of the subsequent stage of parameter estimation. The term 'pseudo-polynomial expressions' is used here because the parameters of any of the expressions in Equation (2) can be computed as in a linear problem and/or as with true polynomial expressions. As mentioned, the parameters a_j are estimated by an LS method integrated into the EPR procedure (Giustolisi & Savic 2006). The LS guarantees a two-way correspondence between the pseudo-polynomial structure and its coefficients. In addition to the usual LS search, the user can force the LS to seek structures that contain only positive coefficients ($a_j > 0$). This is particularly useful in modelling systems where there is a high probability that the negative coefficient values ($a_j < 0$) are selected to balance the particular realization of errors related to the finite training dataset (Giustolisi *et al.* 2007).

Over-fitting in EPR

In regression-based modelling, 'fitness' usually refers to a measure of how closely the regression expression fits the data points. However, it is widely accepted that the best modelling approach is also the simplest which fits the purpose of the application. This principle, often called Occam's razor, is attributed to the medieval philosopher William of Occam (or Ockham, 1300–1349). The so-called principle of parsimony states that for a set of otherwise equivalent models of a given phenomenon one should choose the simplest one to explain a dataset. There is also a need to include a measure of trade-off between model complexity (i.e. the number of parameters) and fitness in regression-based models.

For a given set of data observations, a regression-based technique needs to search among a large, if not infinite, number of possible models to explain those data. By varying the exponents for the columns of matrix $\mathbf{X}$, and by searching for the best-fit parameter set $\boldsymbol{\theta}$, the EPR methodology searches among all those models. It does, however, require an objective function that will ensure the best fit without the introduction of unnecessary complexity. Unnecessary complexity is here defined as the addition of new terms,

or combinations of inputs, that fit mostly random noise in the raw data rather than the underlying phenomenon. The key objective here is therefore to find a systematic means to avoid the problem of over-fitting. In the original single-objective EPR (Giustolisi & Savic 2006) there are three possible approaches to this problem: (1) to penalize the complexity of the expression by minimizing the number of terms; (2) to control the variance of a_j constants (the variance of estimates) with respect to the their values; and (3) to control the variance of polynomial terms with respect to the variance of residuals.

More recently in Giustolisi *et al.* (2007) the idea of using a multi-objective strategy to constrain $a_j > 0$ during parameter estimation for improving model selection (i.e. also as avoidance of over-fitting techniques) was introduced.

Single versus multi-objective GA-based EPR

Although the original EPR methodology proved effective (Giustolisi & Savic 2006), it used only the single-objective genetic algorithm (SOGA) (Goldberg 1989) strategy for exploring the formulae space. In fact, this exploration was achieved by assuming first the maximum number of terms m in the pseudo-polynomial expressions shown in Equation (1) and then sequentially exploring the formulae space having 1, 2, … , m terms. To speed up the convergence, the initial population of each EPR search was (optionally) seeded with the formulae obtained in the previous search (e.g. the population for formulae having j terms was seeded with the best formulae having $j - 1$ terms). However, the SOGA-based EPR methodology has the following drawbacks.

1. Its performance decreases exponentially with an increasing number of polynomial terms m (also because by increasing j, more GA runs are needed).
2. The results are often difficult to interpret. In fact, the set of models identified could be ranked either according to their fitness or according to their structural complexity. However, ranking models according to their structural complexity requires some subjective judgment and, consequently, this process is often biased by the analyst's experience rather than being purely based on mathematical/statistical criteria (Young *et al.* 1996).
3. When searching for the formulae with j terms, those having a smaller number of terms belong to the space of formulae with j terms as a degenerative case. However, these 'degenerative formulae' could have a better accuracy than those previously found (i.e. for lower values of index j) and discarded because at run j there could be less parsimonious formulae that fit the data better.

To overcome these drawbacks, it is possible to use a multi-objective genetic algorithm (Goldberg 1989) (MOGA) strategy in EPR. In fact, assuming m pseudo-polynomial terms (and considering that all pseudo-polynomials having less than m terms belong to the formulae space of m terms as a degenerative case) it is possible to explore the space of m-term formulae using the following two (conflicting) objectives: maximization of model accuracy and minimization of the number of polynomial coefficients in the formulae.

This problem can be solved using the MOGA approach based on the Pareto dominance criterion (Pareto 1896). Adopting this criterion makes the EPR search faster because the search for all models ($j = 1, 2, \ldots, m$) is performed simultaneously. Moreover, the models obtained in this way are already ranked according to: (1) the number of terms obtained (i.e. parsimony) and (2) the accuracy achieved (i.e. model fitness to training data). Following this reasoning, a further improvement to EPR would be to use the MOGA strategy to optimize the number of formulae inputs ($\mathbf{X}_i$ are the vectors). Therefore, objectives of the EPR search are:

1. maximization of the model accuracy;
2. minimization of the number of polynomial coefficients; and
3. minimization of the number of inputs.

Note that EPR can determine the Pareto front consisting of best formulae (maximum of m terms) considering both parsimony (number of constants and variables) and accuracy in a single formula space exploration. This makes EPR results easily interpretable because the formulae are ranked according to the parsimony and accuracy objectives. Moreover, the overall Pareto front gives insight into the model selection phase. Finally, the GA used for the evolutionary stage of EPR is OPTIMOGA. Further details on OPTIMOGA can be found in Giustolisi *et al.* (2004).

Multi-case strategy for EPR

When a given set of observed data is described by different model structures of increasing complexity at least one model structure is returned that allows a correct description of the system in terms of both parsimony and fitness (Ljung 1999). Other structures differ for the selection of variables describing the particular realization of the noise rather than the underlying phenomenon. Such effect goes under the name of over-fitting to training data. Also, polynomial models returned by the EPR usually contain a certain combination of explanatory variables which are common to the majority of Pareto optimal models, whereas other variables or even entire terms are selected in just a few models. In the case of individual pipe systems, the balance among model accuracy, complexity and prior insight into the phenomenon can help in selecting the most suitable model to avoid over-fitting. However, when the same phenomenon is modelled for distinct systems, significant differences may exist among resulting failure models (Berardi *et al.* 2005, 2006). Such observation makes it difficult to separate the description of the underlying physical phenomenon (common to all systems analyzed) from other variables/terms whose relevance emerges from local properties and the particular manifestation of noise in a given measurement of the system. This raises doubts about the correctness of individual system models that are identified and their use as general performance indicators.

Assume that C systems (i.e. cases) $(S1, \ldots, SC)$ exist, each with the relevant observed dataset s $(s = 1, \ldots, C)$ containing data on both recorded sewer failures $\mathbf{Y}_s$ (e.g. collapses or blockages) and the corresponding potential k explanatory variables (i.e. $\mathbf{X}_{s,i}$, $i = 1, \ldots, k$). In such a case, it should be possible to simultaneously identify the best set of k significant explanatory variables for describing the same phenomenon (i.e. $\mathbf{Y}$) in all systems. In MCS-EPR this can be done by first encoding each candidate model structure (as a set of polynomial exponents corresponding to potential explanatory variables in all polynomial terms) and then by using the GA-based EPR search procedure (see previous section) to find the best model structure. During the GA search procedure, each time the potential solutions' fitness is evaluated the following two steps are applied.

(1) Estimate the unknown polynomial coefficient values (i.e. model parameters) by means of numerical regression, such as by using the least squares method. Note that when doing so all model parameters $a_{s,j}$ $(s = 1, \ldots, C,\ j = 1, \ldots, m+1)$ for all individual systems are evaluated simultaneously.

(2) Calculate the three objective function values (sum of squared errors, number of polynomial terms, number of significant explanatory variables; see previous section) to determine each model structure's fitness. Note that the latter two function values do not change from one system to another while the value of the first objective (sum of squared errors) depends on how closely each of the C models (with parameters $a_{s,j}$, $j = 1, \ldots, m+1$) fits its observed data.

It can be argued that there are at least two possible approaches for taking into account different model fitness for the same model structure. The first approach is consistent with the multi-objective paradigm that the MO-EPR is built on and consists of taking each CoD (i.e. the CoD referred to dataset s) as a separate objective to be maximized. The second approach aims to merge all C measures of model accuracy into a single fitness value. The latter approach has been adopted here for the following reasons.

1. A single fitness function value is likely to return a lower number of models than the multi-objective approach which is easier to handle.
2. From a computational point of view, the second approach is faster than the first because of fewer objective evaluations and the reduced number of solutions to be managed.

The following measure of model accuracy is therefore used here:

$$\mathrm{CoD} = 1 - \frac{\sum_{s=1}^{C} \sum_{N_s} (\hat{y}_s - y_{\exp})^2}{\sum_{N} (y_{\exp} - \mathrm{avg}(y_{\exp}))^2} = 1 - \frac{\sum_{s=1}^{C} N_s \cdot \mathrm{SSE}_s}{\sum_{N} (y_{\exp} - \mathrm{avg}(y_{\exp}))^2} \quad (3)$$

where N is the total number of samples over all C datasets (i.e. $N = \sum N_s$); $\mathrm{avg}(y_{\exp})$ is the average value of observations evaluated on N samples; $\hat{y}_s$ is the value predicted by

the model built with the sth vector of parameters and $y_{\exp}$ is the corresponding observation.

APPLICATION TO SEWER FAILURE MODELLING

The MCS-EPR methodology described in the previous section is tested here on a case study consisting of two real UK sewer systems (referred to here as Case 1 and Case 2). Available data consists of: two types of recorded sewer failures (collapses and blockages) recorded during a 5 year monitoring period and pipe data (material, size, age, etc.). Both system datasets contain information on pipes with and without recorded failures. As expected, the number of recorded blockages (2,299 and 2,540 for systems 1 and 2, respectively) largely exceeds the recorded collapses (47 and 37 for systems 1 and 2, respectively).

In addition, all recorded data is only available at the grouped pipe level; that is, 824 (system 1) and 395 (system 2) polygon shape areas (or simply polygons). Each polygon is described by 44 attributes (fields) in the database. The first category of polygon data contains polygon area, number of associated properties, length of main roads and area of 'hazardous' soil (e.g. clay). The second category of polygon data refers to asset features described by the mean sewer age in the area, namely gradient and cover depth which are in turn represented by three sub-classes describing the length of sewers with 'low', 'normal' and 'high' attribute values. For sewer gradients, the classes are 'less than 0.01', 'between 0.01 and 0.05' and 'greater than 0.05'. In the case of cover depth, the relevant thresholds are 0, 1.5 and 3.0 m, respectively.

Sewer nominal diameter is also reported as three sub-classes corresponding to the 'less than 350 mm', 'between 350 and 650 mm' and 'greater than 650 mm'. The third category of polygon attributes refers to asset condition and reports the length of pipes surveyed by means of CCTV, the length of pipes exhibiting the worst operational (ocg) and service (scg) condition grade (e.g. ocg and scg attributes equal to 4 or 5), the length of the so-called 'Section 24' sewers (typically small diameter sewers close to houses) and the length of pipes which experienced surcharges during the monitoring period. Additional information about polygon data available can be found in Berardi *et al.* (2006).

Data pre-processing

As expected, the quality of available data was not ideal. A number of inconsistencies were identified. As a consequence of preliminary analysis, 87 polygons in system 1 and 94 polygons in system 2 were omitted from further analyses.

Selection of potential explanatory variables

Once the cleansing of two datasets was completed, different attributes were enlisted as potential explanatory factors for modelling sewer blockage and collapse. Such a selection was driven by physical insight into the mechanisms leading to different types of failure (collapse and blockage). For modelling blockage, sewer cover depth is neglected in favour of gradient since this is more likely to explain the propensity to obstruction by directly addressing hydraulic conditions. Different mechanisms were identified for modelling sewer collapse. They are caused mainly by the transmission of surface loads and are therefore better explained by cover depth than by sewer gradient. Finally, all other available attributes were chosen as possible explanatory factors for both collapse and blockage.

The EPR function used here is

$$\hat{\mathbf{Y}} = \sum_{j=1}^{m} a_{s,j} (\mathbf{X}_1)^{\mathrm{ES}(j,1)} \ldots (\mathbf{X}_k)^{\mathrm{ES}(j,k)} \tag{4}$$

The maximum number of terms is $m = 3$, and the condition $a_{s,j} > 0$ is used during parameter estimation. The candidate exponents for EPR were $\{-2; -1; 0; 1; 2\}$ in which the choice of 0 allows the procedure to eschew unnecessary inputs.

Results and discussion

Tables 1 and 2 report Pareto optimal one- and two-term polynomial model structures identified by the MSC-EPR for describing the number of collapses (i.e. CL) and blockages (i.e. BL) in systems (i.e. cases) 1 and 2. In addition to this, the CoD value is reported for each model shown. Table 3 outlines a corresponding selection of Pareto optimal models returned by EPR in Berardi *et al.* (2006), where systems 1 and 2 were analyzed individually. Figure 1 displays the reported model structures for blockages and collapses,

Table 1 | Selected Pareto optimal collapse prediction models identified by MCS-EPR

Model structure	a_1—Case 1 a_1—Case 2	a_2—Case 1 a_2—Case 2	CoD—Case 1 CoD—Case 2
$CL = a_1 \cdot \frac{s24^2 \cdot ocg^2}{A^2}$	0.0450	–	0.45
	0.0133	–	0.43
$CL = a_1 \cdot \frac{s24^2 \cdot ocg^2}{A^2} + a_2 \cdot dh$	0.0345	0.0080	0.57
	0.0124	0.0108	0.50
$CL = a_1 \cdot \frac{s24^2 \cdot ocg^2}{A^2} + a_2 \cdot Age \cdot dh$	0.0318	0.0001	0.62
	0.0124	0.0001	0.49

characterized by their performances and selected explanatory variables. The following symbols are used in Tables 1–3: s24 is the length of 'Section 24' sewers; ocgp is the percentage of pipes surveyed by CCTV with the highest (worst) operational condition grade; ocg and scg are the lengths of pipes showing the worst operational and service condition grade; dl and dh denote 'low' and 'high' cover depth classes; Dl and Dm denote 'low' and 'normal' diameter classes; A is the polygon area; Haz is the area of hazard soil in the polygon; s is the length of pipes which experienced surcharge during the monitoring period; and Age is the mean sewer age in the polygon area.

We can note several points from Tables 1–3.

1. Comparison between collapse models shown in Tables 1 and 3 shows a drastic reduction in the number of significant explanatory (i.e. input) variables in the multi-case strategy (see the diagram on the right in Figure 1). This was expected since only a small number of polygons have recorded failures. This fact implies that a higher number of input variables needs to be combined in order to realize an acceptable description of data (e.g. the case for the individual models shown in Table 3).

Table 2 | Selected Pareto optimal blockage prediction models identified by MCS-EPR

Model structure	a_1—Case 1 a_1—Case 2	a_2—Case 1 a_2—Case 2	CoD—Case 1 CoD—Case 2
$BL = a_1 \cdot s24$	9.7435	–	0.86
	17.9908	–	0.69
$BL = a_1 \cdot s24 + a_2 \cdot Haz$	9.6002	22.9537	0.86
	14.9179	28.7993	0.76
$BL = a_1 \cdot s24 + a_2 \cdot Dl^2$	9.7435	0.0000	0.86
	17.4793	0.0082	0.71

Table 3 | Selected Pareto optimal models identified by the EPR technique

	EPR models	CoD
Case 1	$CL = 0.0014268 \cdot \frac{s24 \cdot ocg \cdot dl \cdot dh \cdot}{Dl} + 0.0005888 \cdot Age \cdot s$	0.69
	$BL = 25.1533 \cdot Haz + 0.11091 \cdot Age \cdot s24$	0.90
Case 2	$CL = 2.2355 \cdot 10^{-5} \cdot \frac{Dm \cdot scg^2 \cdot ocgp^2 \cdot s24}{A^2 \cdot ocg} + 0.59796 \cdot Haz$	0.64
	$BL = 15.4937 \cdot s24 + 22.9971 \cdot A$	0.77

2. The same first polynomial term is reported in all three models shown in Table 1. This term contains the following three significant variables: length of Section 24 sewers, length of sewers with the worst operating condition grade and total polygon area. Note that these three significant variables also exist in collapse models for Cases 1 and 2 shown in Table 3 (individual system models). Also, direct/inverse relations between the number of collapses and the previously mentioned three significant explanatory variables are identical in the grouped and individual cases (e.g. reduction in the length of s24 sewers is leading to a reduction in the number of collapses, etc.).
3. Despite their similarities, model structures returned by the MCS-EPR have lower CoD values than the corresponding values obtained in Cases 1 and 2 (see Figure 1, the circle dots against the others). This is a predictable effect of trying to fit the single 'compromise' model to two different observed datasets. As a consequence, failure models generated by the MCS-EPR are not expected to improve individual model fits but rather to provide an 'unbiased' description of the underlying phenomenon by identifying the most significant explanatory variables.

Note that observations similar to the above could also be made in the case of blockage models where the length of Section 24 sewers is identified as the most important explanatory variable. However, unlike in the case of collapse models, the addition of other input variables (e.g. Haz or Dl) improves the model fit in Case 2 only (see blockage models in Tables 2 and 3 and the diagram on the left in Figure 1). This is a consequence of the lower quality of Case 2 data which, incidentally, was recorded by two separate companies during the monitoring period. The presence of unreliable/biased information in the second dataset leads to the selection of variables such as Haz or Dl which are not strictly needed for describing the physical

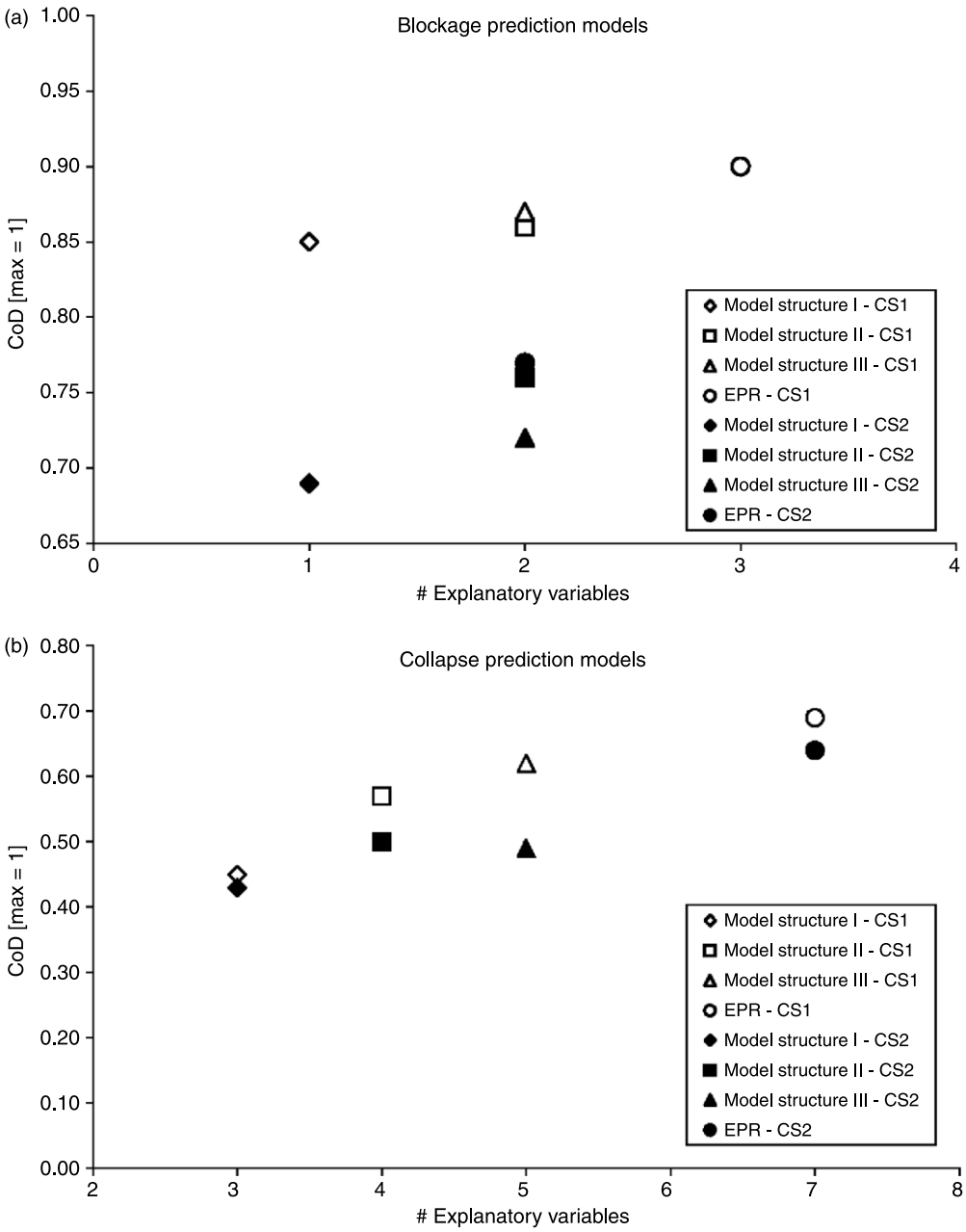

Figure 1 | Performances and explanatory variables of prediction models identified by MCS-EPR and EPR.

phenomenon present in Case 1. Therefore, additional explanatory variables improve the model fit in Case 2 only.

Moreover, unlike the case of collapse models, the number of explanatory variables selected by the MCS-EPR (Table 2) is almost the same as in the EPR case (Table 3), as shown in Figure 1. This is due to the strong imbalance between recorded collapses and blockages. The larger number of blockages, and their distribution among the polygons, allows for a clearer identification of significant explanatory variables. This happens when systems are analyzed both individually (Table 3) and using a multiple case approach (Tables 1 and 2).

Further, note that despite the above differences between individual and multi-case model structures, the variables

selected by the EPR and MCS-EPR strategies, and their influence on failure occurrence (i.e. direct/inverse relations), is still in agreement with prior insight into the underlying system physics. When this is not evident, it should be recalled that the description provided by individual models is affected by local traits or noise in the particular data realization. This is the case of variable ocg in the collapse model reported in Table 3 for Case 2. It is apparent that the length of worst operating condition grade sewers should directly affect collapse occurrence, whereas in that model ocg is located in the ratio denominator. The simultaneous analyses of both cases unearthed a more plausible formulation of the collapse model.

Finally, we note that if different quality data is available for the different cases (i.e. systems) analyzed (which is the case here), this will affect the performance of the corresponding model structures when regressed on (i.e. when model parameters are estimated for) each dataset. In the case study presented here, the quality of Case 2 data was lower than in Case 1. As a consequence, model structures in Tables 1 and 2 return higher CoD values when they are regressed on Case 1 as opposed to Case 2.

APPLICATION TO WATER MAINS DETERIORATION MODELLING

Pipe degradation has commonly been studied as a steady monotonic process that is modified by time-varying 'noise' (Kleiner & Rajani 2002). Time-dependent factors can be random, cyclical (i.e. environmental conditions, traffic loading, external stress, corrosion, etc.) or variable (i.e. operational factors). Pipe age, diameter and material have been identified as primary variables influencing the monotonic increase in the burst rate over a number of years. The majority of statistical methods consider pipe age as the most crucial variable describing the increase in pipe failure rates over time. Exponential (Shamir & Howard 1979) or power (Kleiner & Rajani 2001) models are commonly used to determine the optimal replacement time for pipes. Furthermore, Walski & Pelliccia (1982) found diameter to be a key factor, with the failure rate of smaller diameter pipes being higher than those experienced by larger ones. This is partly due to a potentially lower quality of workmanship associated with laying the pipes (as compared with more expensive larger diameter pipes) and thinner pipe walls. Studies into common metallic pipe behaviour (e.g. cast iron, ductile iron, etc.) have been conducted to establish the influence of pipe material on failure rates (Kettler & Goulter 1985; Kleiner & Rajani 2002).

The study performed on a real network by Pelletier *et al.* (2003) also revealed a close dependence between pipe material, diameter and age. Moreover, age, material and diameter are usually the only, if any, information available to many municipalities and water companies. Long burst data records of high quality and especially at the pipe level are rarely found in practice (Kleiner & Rajani 2002). However, available pipe data together with additional variables such as soil type, land use and/or spatial and temporal clustering of pipe brakes, have been used as grouping criteria to emphasize their influence on failure (Walski & Pelliccia 1982; Kleiner & Rajani 1999; Pelletier *et al.* 2003). Recently, Berardi *et al.* (2005) demonstrated the dependence of pipe bursts on length, age and diameter using real data from UK water companies.

Case study description

This second case study is presented here with two ends in mind: (1) to show the application of MCS-EPR methodology on a larger set of systems and (2) to use MCS-EPR for modelling failure performance indicators of water distribution systems. The database used here refers to 48 water quality zones (WQZ) within a UK water distribution system and consists of about 101,979 items.

The data were available at the pipe level and contain both asset information and bursts recorded during a 14-year monitoring period. For each pipe, the database reports pipe diameter, material, year laid, length, number of properties supplied and the total number of bursts recorded. Unfortunately, neither criteria adopted for designing these water quality zones nor the network map was available for this study. Moreover, the timing of each pipe burst is unknown. These information gaps prevent the introduction of additional variables describing spatial and/or temporal proximity to nearby failures as well as the verification of the potential existence of clusters in the burst data. However, statistical distribution of the asset features (i.e.

Table 4 | Pareto optimal burst prediction models identified by MCS-EPR

	I		II		III	
	$BR = a_1 \cdot Lt^{1.5}$		$BR = a_1 \cdot Ae^{1.5}Lt$		$BR = a_1 \cdot ((Ae^{1.5}Lt)/De)$	
WQZ	a_1	CoD	a_1	CoD	a_1	CoD
1	3.29×10^{-5}	0.5486	7.79×10^{-6}	0.5664	1.12×10^{-3}	0.7958
2	2.57×10^{-5}	0.5056	5.81×10^{-6}	0.5141	6.29×10^{-4}	0.5981
3	2.93×10^{-5}	0.6888	6.75×10^{-6}	0.7273	6.99×10^{-4}	0.7525
4	4.01×10^{-5}	0.6031	9.15×10^{-6}	0.7790	9.12×10^{-4}	0.8165
5	4.01×10^{-5}	0.4849	8.79×10^{-6}	0.7435	8.85×10^{-4}	0.8112
6	1.89×10^{-7}	−0.2213	5.31×10^{-6}	0.3410	5.50×10^{-4}	0.3331
7	3.28×10^{-5}	0.7709	6.46×10^{-6}	0.7268	7.78×10^{-4}	0.8137
8	1.18×10^{-4}	0.6195	1.46×10^{-5}	0.6667	1.24×10^{-3}	0.5777
9	4.96×10^{-5}	0.6805	1.38×10^{-5}	0.7511	1.71×10^{-3}	0.8727
10	7.04×10^{-5}	0.3325	1.08×10^{-5}	0.5381	1.16×10^{-3}	0.6069
11	2.09×10^{-5}	0.7811	4.34×10^{-6}	0.8034	4.41×10^{-4}	0.8349
12	1.67×10^{-5}	0.6622	4.47×10^{-6}	0.8712	4.59×10^{-4}	0.8817
13	5.53×10^{-6}	−0.1132	7.99×10^{-6}	0.6119	8.69×10^{-4}	0.7075
14	4.70×10^{-5}	0.6393	1.10×10^{-5}	0.8012	1.20×10^{-3}	0.7702
15	3.28×10^{-5}	0.9336	5.02×10^{-6}	0.8287	5.08×10^{-4}	0.8385
16	1.83×10^{-5}	0.5956	2.78×10^{-6}	0.6689	3.17×10^{-4}	0.7811
17	1.37×10^{-7}	−0.1124	4.64×10^{-6}	0.5060	5.29×10^{-4}	0.4799
18	8.26×10^{-6}	0.2522	2.48×10^{-6}	0.3892	2.73×10^{-4}	0.4480
19	5.98×10^{-6}	0.2755	2.15×10^{-6}	0.4340	2.46×10^{-4}	0.4359
20	7.04×10^{-6}	0.3367	2.54×10^{-6}	0.4851	2.86×10^{-4}	0.5606
21	7.43×10^{-6}	0.1954	2.50×10^{-6}	0.3468	2.69×10^{-4}	0.3814
22	4.63×10^{-6}	0.4784	1.67×10^{-6}	0.5663	1.86×10^{-4}	0.6518
23	7.65×10^{-6}	0.8925	1.80×10^{-6}	0.8708	1.91×10^{-4}	0.9314
24	1.03×10^{-4}	0.7981	2.17×10^{-5}	0.8732	2.27×10^{-3}	0.9101
25	9.35×10^{-6}	0.5384	3.23×10^{-6}	0.6854	3.29×10^{-4}	0.7150
26	5.20×10^{-5}	0.9270	1.25×10^{-5}	0.9511	1.26×10^{-3}	0.9833
27	5.31×10^{-6}	0.7791	1.18×10^{-6}	0.8132	1.27×10^{-4}	0.8515
28	1.86×10^{-5}	0.8547	4.37×10^{-6}	0.8620	4.50×10^{-4}	0.8551
29	4.69×10^{-5}	0.7949	9.34×10^{-6}	0.8679	1.00×10^{-3}	0.9117
30	1.15×10^{-5}	0.5145	2.88×10^{-6}	0.6336	3.84×10^{-4}	0.8030
31	1.14×10^{-5}	0.3564	5.57×10^{-6}	0.8089	6.08×10^{-4}	0.8792
32	1.88×10^{-5}	0.7543	6.62×10^{-6}	0.8978	6.97×10^{-4}	0.9352
33	2.87×10^{-5}	0.7277	8.56×10^{-6}	0.8690	9.29×10^{-4}	0.9227
34	1.37×10^{-5}	0.6467	2.64×10^{-6}	0.6894	2.90×10^{-4}	0.7234
35	3.70×10^{-5}	0.7495	1.05×10^{-5}	0.8852	1.11×10^{-3}	0.9358
36	3.43×10^{-5}	0.9609	8.47×10^{-6}	0.9395	8.52×10^{-4}	0.9409
37	1.79×10^{-5}	0.7886	5.03×10^{-6}	0.8594	4.72×10^{-4}	0.8861
38	2.83×10^{-5}	0.9028	7.02×10^{-6}	0.8902	5.74×10^{-4}	0.9414
39	1.57×10^{-5}	0.8674	5.02×10^{-6}	0.9101	4.03×10^{-4}	0.9433

Table 4 | (*continued*)

	I		II		III	
	$BR = a_1 \cdot Lt^{1.5}$		$BR = a_1 \cdot Ae^{1.5}Lt$		$BR = a_1 \cdot ((Ae^{1.5}Lt)/De)$	
WQZ	a_1	CoD	a_1	CoD	a_1	CoD
40	1.34×10^{-5}	0.9097	4.31×10^{-6}	0.9086	4.35×10^{-4}	0.9194
41	1.77×10^{-5}	0.7583	3.02×10^{-6}	0.6752	2.33×10^{-4}	0.5867
42	2.52×10^{-5}	0.9339	5.27×10^{-6}	0.9311	4.49×10^{-4}	0.9468
43	2.36×10^{-5}	0.7970	6.02×10^{-6}	0.8771	5.82×10^{-4}	0.9463
44	3.74×10^{-5}	0.6434	3.80×10^{-6}	0.7370	3.49×10^{-4}	0.7900
45	2.16×10^{-5}	0.4559	1.24×10^{-5}	0.4662	1.43×10^{-3}	0.4799
46	1.68×10^{-5}	0.7851	3.85×10^{-6}	0.8129	3.78×10^{-4}	0.8126
47	7.93×10^{-6}	0.5053	2.65×10^{-6}	0.7203	2.69×10^{-4}	0.7244
48	6.42×10^{-6}	0.7354	2.25×10^{-6}	0.7900	2.78×10^{-4}	0.8839

diameter, vintage, total length of pipeline and total number of properties supplied) allows all the WQZs to be considered as stand-alone small water distribution systems.

As in the majority of water distribution systems, the number of failures recorded during the monitoring period corresponds to about 10% of the total number of pipes and several pipes failed more than once over the same time period. A performance indicator should represent the propensity to fail for all pipes in the network. Therefore, both pipes with and without recorded bursts have been considered. Furthermore, the distribution of bursts mentioned above implies a grouping criterion to be adopted in order to have a finite failure rate for all pipes in the network. Previously developed pipe failure models (Shamir & Howard 1979; Kleiner & Rajani 1999; Giustolisi & Savic 2004) associated the same pipe failure rate to pipes with similar attributes (e.g. material, size, age, etc.). Following on from that work, and based on the preliminary analyses, the pipes considered here have been classified using pipe diameter and age.

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. These pipes have been grouped into 8 diameter classes (from 32 mm to 250 mm). Such a classification has been used to fill in some existing data gaps. In fact, numerous records contained 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. Within each diameter class, the mean burial year of pipes made of the same material was used to complete missing data. Once the data reconstruction was completed, pipes were further grouped into 1-year age classes. The choice of 1-year for age classification is because it averages the influence of time-dependent factors over a year and allows for detailed analysis of the problem based on data updated annually by water utilities.

Selection of potential explanatory variables

Only four fields describing pipe features have been considered for modelling. These are age, diameter, length and number of properties supplied, all available at the pipe level. For each diameter-age class, the total number of recorded burst events (BR), the sum of pipe lengths (Lt), the sum of properties supplied (Pr) and the total number of pipes in the class (Np) have been computed. In summary, the model under consideration is geared to identify the functional relationships between five possible model inputs (Ae, De, Lt, Np, Pr) and one model output (BR). The same EPR model structure reported in Equation (4) was used in this case study except for the maximum number of terms allowed (which is $m = 1$ here) and the set of exponents $\{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$.

Results and discussion

Table 4 reports optimal model structures obtained by MCS-EPR for describing pipe burst occurrence in a water

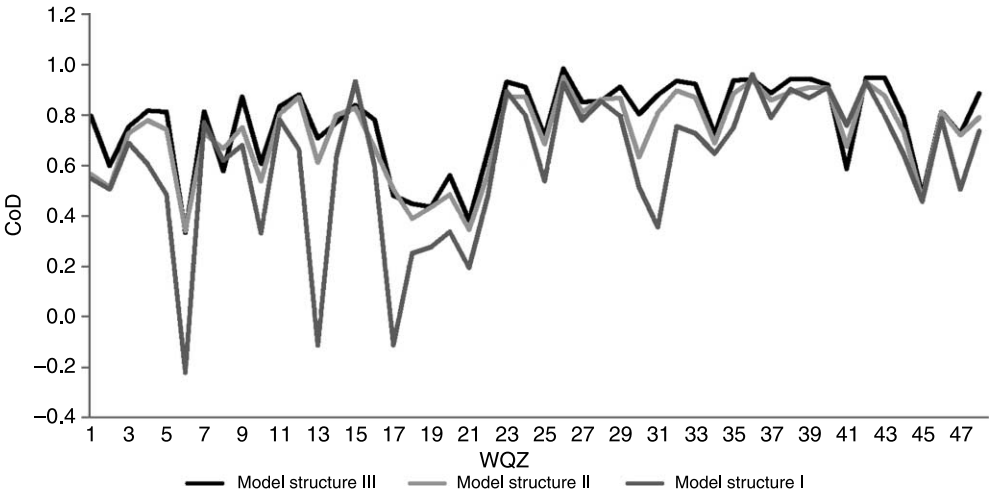

Figure 2 | Performance of burst prediction models identified by MCS-EPR.

distribution network. For each zone the CoD values and coefficients (a_1) are reported. A pictorial representation of model performance is reported in Figure 2 which shows that the addition of one or two input variables improves performance for almost all WQZs. In particular, model structures I, containing total class length Lt, leads to 18 cases described with CoD less than 0.60 and 3 cases with negative CoD values. Note that negative CoD means that the average observation (here total number of bursts in the classes) would provide a better description than model structure I.

The selection of the overall class length Lt 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.

Addition of the age term Ae leads to an average increase of CoD of about 0.116 and a significant improvement of performance in almost all cases. Direct dependence on age confirms this variable to be the most significant factor in describing the deterioration process and subsequent burst occurrence in a water distribution network. It is noteworthy that in this case study the variable Ae also includes information on pipe material since it has been used for infilling missing data on age.

Model structure III is the most complex returned by MCS-EPR and contains just one more explanatory variable (i.e. equivalent diameter De). From models II to III there is an average increase of CoD of about 0.42 and the system description has improved for 41 cases. Also in this case the inverse dependence between pipe diameter and number of bursts occurring in the network confirms the observation that smaller pipes are more prone to failure than larger ones (Clark *et al.* 1982; Walski & Pelliccia 1982; Kettler & Goulter 1985).

CONCLUSIONS

A novel approach for generating polynomial-type sewer failure prediction models from observed data is developed and presented here. The MCS-EPR approach uses the existing EPR technique to simultaneously identify the best model structure and parameter values from the observed data available for multi-utility data (i.e. cases). This way, the resulting models for predicting the number of pipe failures should contain only the explanatory factors important for describing the underlying physical phenomenon. The advantages of using such an approach are: (1) the approach results in a generalization of the EPR outcomes and the formulation of more realistic failure models; and

(2) the approach itself provides the mechanism for the selection of the most general model structures without asking the analysts for a comparison between models returned for different systems.

The MCS-EPR methodology presented here has been tested and verified on both real drinking and wastewater systems in the UK for predicting the number of pipe bursts, collapses and blockages. The results obtained show several advantages of this approach when compared to the existing EPR approach applied to individual systems. The main advantages are: (1) models returned by the MCS-EPR provide and verify the physical insight into the underlying physical phenomenon (pipe failure); (2) incorrect relationships between the number of failures and some explanatory variables and/or physical misinformation achieved by the individual system level EPR analysis can be identified and overcome; and (3) varying quality of different datasets corresponding to different sewer systems is stressed by different performances when a given MCS-EPR model structure is applied to each single system.

ACKNOWLEDGEMENTS

The first author gratefully acknowledges the support of the UK Engineering and Physical Sciences Research Council (Platform grant, GR/T26054). The authors are also grateful to L. Berardi for his ideas and numerous fruitful discussions on the topic of this paper.

REFERENCES

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

Ariaratnam, S. T., El-Assaly, A. & Yang, Y. 2001 Assessment of infrastructure inspection needs using logistic models. *J. Infrastruct. Syst.* **7** (4), 160–165.

Babovic, V. & Keijzer, M. 2000 Genetic programming as a model induction engine. *J. Hydroinform.* **2** (1), 35–61.

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

Berardi, L., Savic, D. A. & Giustolisi, O. 2005 Investigation of burst-prediction formulas for water distribution systems by evolutionary computing. In *Proceedings of the 8th International Conference on Computing and Control for the Water Industry, Exeter, UK*, (Vol. 2), pp. 275–280, Centre for Water Systems, Exeter, UK.

Berardi, L., Kapelan, Z., Savic, D. A. & Giustolisi, O. 2006 Modelling sewer performance indicators. In *Proceedings of Hydroinformatics 2006, Nice, France*, (Vol. 4), 2829–2836, Research Publishing, Chennai (India).

Cardoso, M. A., Coelho, S. T., Matos, J. S. & Matos, R. M. 1999 A new approach for diagnosis and rehabilitation of sewerage systems through the development of performance indicators. In *Proceedings of 8th International Conference on Urban Storm Drainage, Sydney, Australia*, (Vol. 2), pp. 610–617, IAHR Publishing, London.

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

Draper, N. R. & Smith, H. 1998 *Applied Regression Analysis*. John Wiley and Sons, New York, USA.

Giustolisi, O. & Savic, D. A. 2004 Decision support for water distribution system rehabilitation using evolutionary computing. In *Proceedings of the Seminar on Decision Support in the Water Industry under Conditions of Uncertainty (ACTUI), Exeter, UK*, pp. 76–83, Exeter University Press, Exeter, UK.

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

Giustolisi, O., Doglioni, A., Laucelli, D. & Savic, D. A. 2004 *A proposal for an effective multiobjective non-dominated genetic algorithm: the OPTimised Multi-Objective Genetic Algorithm, OPTIMOGA*. Report 2004/07, School of Engineering Computer Science and Mathematics, Centre for Water Systems, University of Exeter, UK.

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

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

Goldberg, D. E. 1989 *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley. London, UK.

Holland, J. 1975 *Adaptation in Natural and Artificial Systems*. The University of Michigan Press. Ann Arbor, Michigan, USA.

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

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

Kleiner, Y. & Rajani, B. B. 2001 Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water* **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 Means of Natural Selection*. MIT Press, Cambridge, MA, USA.

Ljung, L. 1999 *System Identification: Theory for the User*, 2nd edition. Prentice-Hall Inc. Englewood Cliffs, New Jersey, USA.

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, UK.

Pareto, V. 1896 *Cours D'Economie Politique*, Vol. I and II. Rouge and Cic, Lausanne, Switzerland.

Pelletier, G., Mailhot, A. & Villeneuve, J. P. 2003 Modelling water pipe breaks–three case studies. *J. Water Resour. Plann. Manage.* **129** (2), 115–123.

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

Shamir, U. & Howard, C. D. D. 1979 An analytic approach to scheduling pipe replacement. *J. Am. Water Works Assoc.* **71** (5), 248–258.

Tran, D. H., Nga, A. W. M. & Perera, B. J. C. 2007 Neural networks deterioration models for serviceability condition of buried stormwater pipes. *Eng. Appl. Artif. Intell.* **20** (8), 1144–1151.

Walski, T. M. & Pelliccia, A. 1982 Economic analysis of water main brakes. *J. Am. Water Works Assoc.* **74** (3), 140–147.

Young, P., Parkinson, S. & Lees, M. 1996 Simplicity out of complexity in environmental modelling: Occam's razor revisited. *J. Appl. Stat.* **23** (2–3), 165–210.

First received 12 March 2008; accepted in revised form 9 February 2009