Detecting anomalies in water distribution networks using EPR modelling paradigm

Daniele Laucelli, Michele Romano, Dragan Savić and Orazio Giustolisi

ABSTRACT

Sustainable management of water distribution networks (WDNs) requires effective exploitation of available data from pressure/flow devices. Water companies collect a large amount of such data, which need to be managed correctly and analysed effectively using appropriate techniques. Furthermore, water companies need to balance the data gathering and handling costs with the benefits of extracting useful information. Recent approaches implementing data mining techniques for analysing pressure/flow data appear very promising, because they can automate mundane tasks involved in data analysis process and efficiently deal with sensor data collected. Furthermore, they rely on empirical observations of a WDN behaviour over time, allowing reproducing/predicting possible future behaviour of the network. This paper investigates the effectiveness of the evolutionary polynomial regression (EPR) paradigm to reproduce the behaviour of a WDN using online data recorded by low-cost pressure/flow devices. Using data from a real district metered area, the case study presented shows that by using the EPR paradigm a model can be built which enables the accurate reproduction and prediction of the WDN behaviour over time and detection of flow anomalies due to possible unreported bursts or unknown increase of water withdrawal. Such an EPR model might be integrated into an early warning system to raise alarms when anomalies are detected.

Key words | data mining, evolutionary polynomial regression, timely burst detection, unreported bursts, water distribution networks

INTRODUCTION

Sustainable management of water distribution networks (WDNs) requires the reduction of water leakages from pipelines. This can diminish the waste of a precious resource, decrease the costs of treatment and pumping, minimise third-party damages and, ultimately, lessen greenhouse gas emissions. To this end, the timely detection and location of pipe bursts in a WDN is of fundamental importance. Pipe bursts represent a potential risk to public health and can cause significant environmental damage and economic loss, especially when they remain hidden (i.e., unreported bursts). Burst duration can be divided conceptually in awareness, location and repair time. Often, there is a gap between a burst occurring and the water utility becoming aware of it. Indeed, water companies generally become aware of a burst occurrence through customer contact (e.g., complaints for low pressure, discolouration, etc., or when signs of visible surface water appear). The fact that large bursts usually are rapidly fixed due to multiple complaints, other bursts that do not result in significant impacts on the water delivery service can run undetected for long periods, thus leading to higher overall water losses (WRc 1994).

Among conventional solutions to burst detection and location, water utilities currently use flow-monitoring techniques to assess leakages coupled with operative methodologies. Flow-monitoring techniques are generally
off-line through the application of mass balance type calculations or through observations of changes in specific nighttime values (i.e., minimum night flow (MNF) analysis) (Puust et al. 2010). Operative methodologies usually employ highly specialised hardware equipment, such as leak-noise correlators (Grumwell & Ratcliffe 1988) and pig-mounted acoustic sensors (Mergelas & Henrich 2005). By monitoring MNF, unusual changes in water volumes can be detected. Therefore, the identification and quantification of water losses rely on accurate estimation of expected night flows (McKenzie & Seago 2005) and the MNF data analysis has tended to be a manual or semi-manual process with inherent inefficiencies and prone to human error. Furthermore, the MNF data analysis does not generally look at individual time series for short-term events and averaging over time is usually used (Wu et al. 2011). To summarise, although this technique can be effective under certain circumstances, the MNF data analysis has several limitations and it is not necessarily conducted regularly due to its heavy reliance on manual processes and subjective interpretation of the results obtained. On the other hand, highly specialised hardware equipment is generally used as part of a leak detection survey (Covas et al. 2008) in which temporary zoning may be undertaken. Such an approach can be expensive and time-consuming and even require the shutdown of pipeline operations for long periods.

Another largely used approach to detect anomalies in WDNs is based on the setting of flat-line alarm levels at key monitoring locations in a WDN, allowing near real-time identification of, usually, large bursts. The alarm level values are set as the average of the daily high and low values observed over a 12-month period, plus or minus a certain percentage (i.e., confidence factor). By using this approach, mainly due to spurious measurements, a large number of alerts may be raised by the flat-line alarm system, a significant number being ghosts (a false alarm that no known events correlates with), and yet many events are not detected prior to customer contacts. For this reason, a significant issue in setting flat-line alarm levels is the trade-off between ghosts and non-detection of smaller events (Mounce et al. 2010).

The latest developments in hydraulic sensor technology and on-line data acquisition systems have enabled water companies to deploy a large number of pressure and flow devices. Data collected by these devices provide a potentially useful source of information for reproducing and predicting the behaviour of a WDN. These data are in the form of time series (i.e., a data stream consisting of one or more variables whose value is a function of time) and, when used in conjunction with reproductions/predictions of the WDN behaviour, have the potential to enable fast and economic detection and location of pipe bursts (Romano et al. 2014a, 2014b). In view of this and of the limitations of the aforementioned conventional solutions to the burst detection and location problem, it is clear that new and more efficient techniques are needed for efficiently and effectively exploiting the water industry’s pressure and flow data.

Some techniques use hydraulic simulation models with bursts detected by correlating measurements to expected hydraulic network model’s results via, for example, genetic algorithms (e.g., Wu et al. 2010). Other approaches include the use of inverse transient analytic methods (e.g., Covas et al. 2005), or tackling both the calibration of mathematical simulation tools and the burst detection and location problems at the same time (e.g., Kapelan et al. 2000; Puust et al. 2006).

Another category is that comprising pressure analysis approaches that work by monitoring pressures (or flow) at a number of locations in the network and searching for deviations from the normal/expected pressure trends caused by occurring bursts (Poulakis et al. 2003; Misunias et al. 2005; Shinozuka & Liang 2005; Yamamoto et al. 2006).

The aforementioned techniques have all shown little practical applicability to real-life situations so far. Data-driven approaches comprising soft computing and machine learning (i.e., artificial intelligence) techniques that automatically exploit the constant flow of data (i.e., time-series) coming from monitoring systems to detect unusual changes in the process variable patterns, are found to be the most promising techniques in the context of on-line burst detection and location in real-life WDNs. These techniques are capable of self-learning with a limited number of patterns and able to deal with often patchy, poor-quality data. They enable the extraction of useful information (required for making reliable operational decisions) from the vast and often imperfect sensor data collected by modern Supervisory Control And Data Acquisition (SCADA) systems. They do not need
expensive high frequency measurements, require limited manual or semi-manual interpretation of the results obtained and, most importantly, they only rely on the empirical observation of a WDN behaviour over time. That is to say, they do not need detailed knowledge of the pipe network (e.g., asset parameters). These kinds of approaches are based on a common framework, such as: (i) data preparation (e.g., denoising, data reconstruction); (ii) prediction of expected values based on data-driven models; and (iii) identification of anomalies in flow/pressure signals and raising alerts based on a mismatch between model predictions and signals from meters (Berardi et al. 2014).

An example of data-driven techniques applied to the leakage detection and location problem are the artificial neural networks (ANNs) combined with fuzzy logic technology (Mounce et al. 2006, 2010; Mounce & Boxall 2010), or as self-organising map (Aksela et al. 2009), or combined with a probabilistic inference engine based on Bayesian networks (Romano et al. 2014a, 2014b). Other examples are the application of Kalman filtering techniques (Jarret et al. 2006; Ye & Fenner 2011), support vector machines (Mounce et al. 2011), geostatistical techniques (Kriging) (Romano et al. 2013) and comparison of flow pattern distributions (Van Thienen 2013).

The data-driven approaches mentioned above have been used for burst detection and location with varying degrees of success and different limitations.

The present paper investigates the potential of the evolutionary polynomial regression (EPR) modelling paradigm for burst/other-events detection. The aim of this study is to demonstrate that the EPR modelling paradigm can be used to reliably highlight possible problems in a WDN using pressure/flow measurements at a few points in the network.

The EPR allows the exploration of polynomial models, based on a multi-objective optimisation scheme, where candidate optimals are included in a Pareto set of solutions based on the accuracy of predictions and parsimony of the symbolic model expressions. Therefore, the EPR modelling paradigm presents some beneficial features, not found in other data-driven techniques, such as:

(i) a small number of parameters to be estimated (i.e., helps avoiding over-fitting problems, especially for small data sets);

(ii) a linear parameter estimation methodology that assures the unique solution is found when the inverse problem is well-conditioned;

(iii) automatic model construction (avoiding the need to preselect the functional form and the number of parameters in the model);

(iv) a transparent (‘white box’) form of the regression characteristics, which makes model selection easier, i.e., the multi-objective feature allows selection not only based on fitting statistics.

Consequently, once the Pareto set of optimal models is obtained, the analyst can evaluate the models considering also the key aspects not encoded as objective functions, as for example:

(i) the model structure with respect to physical insights related to the problem;

(ii) similarities of mathematical structures among EPR Pareto set of models;

(iii) recurrent groups of variables in different EPR models;

(iv) generalisation performance of models as assessed in terms of both statistical analysis and mathematical parsimony;

(v) reliability of experimental data used to build the model and/or final purpose of the model itself. This eventually results in a more robust selection.

A number of specific applications of EPR in civil and environmental engineering can be found, for example in Laucelli & Giustolisi (2011), where EPR is used to predict scour depth downstream of grade-control structures, in El-Baroudy et al. (2010) for modelling of the evapotranspiration process, in Laucelli et al. (2014) for bursts modelling based on climate variables, or for prediction of torsional strength of beams in Fiore et al. (2012). In the general framework described above, EPR could be effectively used to improve and/or complement other burst detection and location approaches.

The idea is to use the multi-case (MCS) EPR strategy (Berardi & Kapelan 2007; Savić et al. 2009; Berardi et al. 2014) to develop a water consumption prediction model using values recorded over a number of past weeks (i.e., time windows) that are treated as separate data sets. This involves the search for the same mathematical structure
(i.e., formula) shared by several prediction models, with different sets of model parameters, each minimising the error over a different weekly data set. As explained in the following sections, this returns a range of predictions for the WDN’s water consumption given the pressure/flow measurements at few points in the network. In this context, the measured data are compared to MCS EPR models prediction range, which takes into account the history of the water consumption habits of the customers, to detect abnormal/unexpected changes. Historical records of customer contacts (i.e., claims) following water service problems and records of pipe repair interventions carried out by the utility are then used to evaluate the results obtained, together with an analysis of the MNF variations registered after the detection events.

METHODOLOGY

Modelling approach: multi-case EPR

EPR is a hybrid modelling technique that allows the exploration of polynomial models, where candidate inputs are included in the final model based on the accuracy of predictions and parsimony of the symbolic model expression (Giustolisi & Savic’ 2006). A pseudo-polynomial structure for model expression is used, where each term comprises a combination of candidate inputs (i.e., variables). Each variable gets its own exponent to be determined during the evolutionary search and each polynomial term is multiplied by a constant coefficient, which is estimated by minimising the error on training data. Each monomial term can include user-selected functions among a set of possible alternatives.

An example of the general model structures that EPR can manage is reported in Equation (1), although the interested reader can find more details about the EPR paradigm in Giustolisi & Savic’ (2006):

\[ Y = a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \cdots (X_k)^{ES(j,k)} \]

\[ f \left( (X_1)^{ES(j+1,k)} \cdots (X_k)^{ES(j,2k)} \right) \quad (1) \]

where \( m \) is the number of additive terms, \( a_j \) are numerical parameters to be estimated, \( X_i \) are candidate explanatory variables, \( ES(j,z) \) (with \( z = 1, \ldots, 2k \)) is the exponent of the \( z \)th input within the \( j \)th term in Equation (1), \( f \) is a user-selected function among a set of possible alternatives (including no function selection). The exponents \( ES(j,z) \) are selected from a user-defined set of candidate values (which should include 0).

The search problem is defined in a multi-objective optimisation context, where candidate models are evaluated based on three criteria, namely: (1) model accuracy (maximisation of fitness to data); (2) parsimony of covariates (minimising the number of explanatory variables included in the final model expressions); and (3) parsimony of mathematical equation (minimisation of the number of polynomial terms). The role of the parsimony criteria in EPR is to prevent over-fitting the model to the data, and thus to endeavour capturing the underlying general phenomena without replicating the noise in the data. In this way, the technique can identify the most important input variables for the phenomena under study. The EPR modelling technique uses a multi-objective genetic algorithm (MOGA) optimiser to find candidate models and rank them utilising the above-mentioned criteria and the Pareto dominance methodology (Giustolisi & Savic 2009).

During the evolutionary search, the exponents are selected from a user-defined set of candidate values, which usually include a zero value as well, i.e., an input variable raised to the power of zero is de facto excluded from the model (Giustolisi & Savic’ 2006). At each generation, all the candidate models have a different number of terms and combination of inputs. The constant coefficients are estimated using the available training set, and then the candidate models are selected based on a multi-objective scheme.

Once the symbolic model expressions are obtained, their preliminary validation is based on the physical knowledge of the phenomena being analysed. In addition, the recurrent presence of certain input variables in several non-dominated models indicates the robustness of these inputs as potential explanatory variables of the phenomena. All these features make the EPR modelling paradigm substantially different from purely regressive methods (e.g., ANNs), where statistical measures of accuracy of model predictions are the only criterion that drives model selection,
while final mathematical expressions can be rarely validated from a physical perspective (Savić et al. 2009).

When the available data refer to different realisations of certain physical phenomena under various conditions/observations, it can be more difficult to identify the pattern among variables describing the underlying (i.e., main) system behaviour (Berardi & Kapelan 2007). The MCS EPR is suitable for situations where data can be partitioned into subsets, each representing a particular realisation/experiment of the same phenomena. Thus, the MCS EPR simultaneously identifies the best pattern among significant explanatory variables describing the same phenomena in all data partitions, while neutralising possible impacts of errors and uncertainty in the data. The MCS EPR also makes use of the MOGA optimisation scheme, as described above, where each candidate model structure is evaluated on each considered data partition (Savić et al. 2009).

In particular, the MCS EPR accounts for the model accuracy using the following indicator (coefficient of determination (CoD)):

$$CoD = 1 - \frac{\sum_{i=1}^{C} N_i \sum_{j=1}^{N_i} (\hat{y}_{ij} - y_{ij})^2}{\sum_{i=1}^{C} N_i \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_i)^2}$$

where $C$ is the number of data subsets; $N_i$ is the number of time steps in data subset $i$; $y_{ij}$ and $\hat{y}_{ij}$ are the observed and predicted output values, respectively, in time step $j$ of subset $i$; $\bar{y}_i$ is the average observed output in subset $i$; and $SSE_i$ is the sum of squared errors for data subset $i$ (Savić et al. 2009).

**Modelling procedure description**

The aim of the MCS EPR modelling procedure described here is to reproduce the behaviour of a WDN or a portion of a WDN (e.g., a district metered area, DMA) using observed (and possibly cheap) measurements of hydraulic variables such as pressure and flow at a few points in the network. The behaviour of a WDN can be represented in terms of average flow rate $\Delta H(t)$ or volume of delivered water during the measurement step (e.g., 15 minutes, 1 hour, etc.). The procedure can be applied to normal WDN operating conditions. If the observed values of water consumption are outside the range predicted by the model, the WDN is assumed to be experiencing an anomaly (as it deviates from the expected behaviour). Pre-processing (e.g., filtering, denoising, etc.) can be performed on the available data if the presence of errors can be hypothesised. This said, however, the previously mentioned abilities of MCS EPR allow preventing over-fitting to training data without noise replication (Savić et al. 2009).

Bearing in mind the above, in the MCS EPR modelling procedure described here, the available data sets are divided into a number of weekly data sets ($S + T$ weeks). Two different MCS EPR modelling threads are performed to account for the differences between working days and weekends (i.e., Saturdays and Sundays – during which it is assumed that the customers’ habits are different from the other days in the week). Please note that any other time aggregation could be used to account for the peculiarities of the phenomena under scrutiny (e.g., daily, monthly variations).

The first $S$ weeks are used as training data set to build a set of MCS EPR models large enough to provide a meaningful range of predictions (i.e., historical behaviour of the network). After the initial training phase, the next $T$ weeks are considered for an iterative training/testing phase of the models selected among those returned by the MCS EPR procedure. In particular, with reference to Figure 1, week $T_1$ is first used to test the model trained on the first $S$ weeks; in the next test sub-phase, the MCS EPR models are trained on the previous $S + 1$ weeks (i.e., week $T_1$ is now included in the training data set), using week $T_2$ as test week; and so on, until week $T_{T-1}$ (which is the last used for both sub-phases).

At each recalculation, among the various MCS EPR models returned by the procedure (i.e., Pareto front), the preferred model is selected by trying, where possible, to confirm the structure of the model used in the previous testing sub-phases.

The choice to perform the MCS EPR modelling procedure, instead of only re-calibrating the model’s parameters based on the new training data set, is motivated by the aim to identify the main functional relationships among inputs and output data. This is possible because EPR (due to its aforementioned features) is able to evolve common prediction model’s mathematical structures, as...
representative of the underlying phenomena, for all the considered data sets.

For each weekly data set in the training set, a string of \( m \) parameters \( (a_1 \ldots a_m) \) are determined, where \( m \) is the number of additive terms of the selected MCS EPR model (see Equation (1)). These parameters are representative of the water consumption history of the WDN during the considered week. Every data set has different model parameters and the same model structure; this leads to a range of predictions for the water consumption, one for each analysed data set. The range of predictions reflects different past customers’ behaviour (i.e., weekly demand and related pressure patterns) (Berardi et al. 2014).

As a clarifying example, considering week \( T_1 \), by using the measured inputs (i.e., of flow and pressure) at each sampling time \( t \), the preferred MCS EPR model produces \( S \) predictions of \( \Delta F(t) \) (i.e., water consumption at time \( t \)) given \( S \) different strings of parameters \( (a_1 \ldots a_m) \), one for each of the \( S \) training weeks. In view of this, the value \( \Delta F^{\text{mean}} \), calculated as the average of the \( S \) predictions, can be considered as the most probable expected water consumption at time \( t \) given the inputs included in the selected model. If the observed value of \( \Delta F(t) \) is below \( \Delta F^{\text{mean}} \), this means that at time \( t \) customers are withdrawing less water than usual and this does not result in an alarm. On the contrary, when the observed value of \( \Delta F(t) \) is above \( \Delta F^{\text{mean}} \), customers are withdrawing more water than expected, with respect to the past history of the WDN. This condition can indicate a possible significant change in the behaviour of the system when the observed value of water consumption is approaching the predicted maximum value, \( \Delta F^{\text{max}} \). The way in which the exceedance of \( \Delta F^{\text{mean}} \) (or the proximity to \( \Delta F^{\text{max}} \)) defines an anomaly depends on requirements and experience of the water utility operators, as well as on the network history. For example, if the training weeks are very similar to each other (i.e., the weekly consumptions have comparable values hour by hour and no big leakage of water occurred), the maximum value \( \Delta F^{\text{max}} \) is a threshold easily surmountable by the observed values if a significant leakage (much greater than the minor ones that occurred during the training weeks) occurs during the testing week. In contrast, if among the training weeks there is one or more weeks experiencing a big water loss, there will be more variability in the model predictions and thus the maximum value \( \Delta F^{\text{max}} \) is representative of a really extreme situation occurring in the network life.

A continuous sequence of anomalies, with eventual exceedance of the alarm threshold might then indicate the

---

**Figure 1** | Schematic of the iterative training/testing procedure proposed.
presence of a significant problem that needs to be addressed promptly, i.e., an intervention, as described in the following application.

**CASE STUDY**

**Case study description**

The application used a database coming from the monitoring of pressure/flow in a DMA during normal operating conditions. The data come from three locations in the DMA: the inlet point, the outlet point and a critical pressure point within the DMA. The available database can be considered as a time series, with a measurement time step of 15 minutes, thus a full day observation consists of 96 records. A total of 21 weeks of data are available for this study. No pre-processing (filtering, denoising, etc.) has been performed on the available data, except for data gaps’ removal. The used data have been divided into a number of weekly data sets, differentiating between the weekend (Saturday and Sunday) and the working days of the week, because during the weekend the morning peak is postponed with respect to working days (see Figure 2). Thus, two different modelling procedures are considered. The procedures include a training phase using the first $S = 11$ available weekly data (working days and weekend days), and an iterative testing phase exploiting the last $T = 10$ available weeks.

The used database starts from 6th June 2008 and ends 16th November 2008. The first 11 weeks go from 16th June to 29th August 2008, and the last 10 weeks from 8th September to 16th November 2008. The available inputs are the pressure ($P_1$), flow ($F_1$) measured at gauge N1, the pressure ($P_3$) at gauge N3, and the pressure ($P_2$) and flow ($F_2$) measured at gauge N2. The number of pressure

![Figure 2](https://iwaponline.com/jh/article-pdf/18/3/409/478927/jh0180409.pdf)
observations is quite low; this does not facilitate the reading of the pressure drops in the network when a break occurs.

Therefore, in order to represent the average behaviour of pressure during the analysed periods and to generalise the present application, the average pressure (PM) among the three available measurements has been used in this application. The output data ($\Delta F$) consists of a time series of water consumption (litres per second) calculated as the difference between the water flow at the inlet point ($F_1$) and the water flow at the outlet point ($F_2$) of the DMA; thus, the water consumption is here considered as an average flow over the considered time step, instead of water volume.

Figure 3 shows the DMA layout, the location of measurement points and the elevation of nodes in the analysed DMA. The DMA has a total mains length of 24 km, with 16 boundary valves and no pressure reducing valves. The area contains 2,640 domestic properties and 500 commercial properties of which 48 have a demand greater than 400 m$^3$/year. The zone is predominantly urban domestic/industrial. A number of customer contacts and intervention of the utility crews are available for the period analysed, as reported in Table 1. These data will be useful for identifying/confirming anomalies during the testing phase.

RESULTS AND DISCUSSION

For MCS EPR runs, the candidate set chosen for the exponents were $[-4, -3.5, -3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4]$. This wide range was selected in order to allow the possible inclusion of well-known relationships, e.g., linear, quadratic, inverse linear, square root, powers, etc., for each term involved in the models. Such a modelling choice was mainly aimed towards finding formulations that are as general as possible. The number of past time steps considered for input data was set to 4 (i.e., 1 hour), while no past values of output data were used. The number of polynomial terms was set to $m = 3$, without any constant value in the polynomial expression (bias). The MOGA process was set to run for 5,000 generations. The objective functions are: (1) maximisation of model accuracy as measured by the CoD (Giustolisi & Savic 2006); (2) minimisation of the number of explanatory variables; and (3) minimisation of the number of polynomial terms.

Each EPR model’s search procedure terminated with a Pareto front of optimal solution according to the above-mentioned objective functions. Throughout the entire procedure, checks were made to identify any repeating model structures, thus trying to understand a possible clear relationship between the available input and output. After the training phase, almost all the EPR returned models had a very high accuracy to
Given the predictions of \( \Delta F(t) \) (i.e., \( S \) for week \( T_1, S + 1 \) for week \( T_2 \), etc.), since they refer to a limited period of the network consumption history, it is reasonable to assume a certain probability density function to statistically represent the possible behaviour of the network. As a first attempt, this study assumes that such predictions are characterised by a Normal distribution with average \( \Delta F_{\text{mean}} \) and a standard deviation \( \sigma \) (calculated on the available predictions). This is mainly driven by the aim to give more weight to the average value of predictions with respect to the extreme values, while using a well-known probability density function. Future studies will be also focused on exploiting the potentialities of alternative probability density functions.

This said, it is possible to calculate the cumulative probability of every value of the observed \( \Delta F(t) \). If this cumulative probability is higher than a threshold value, the observed \( \Delta F(t) \) can be considered as a possible anomaly. According to the aforementioned considerations on network history influence on such threshold value (see the previous section), it has been observed that the used training weeks have very similar consumption patterns. In particular, considering all the standard deviations calculated on the available predictions for each time step (hour) in the training set, they have an average value of 0.182 L/s and a standard deviation \( \sigma = 0.031 \); this means that, on average, for each step considered in the training set, all the predictions of \( \Delta F(t) \) are very close to their value of \( \Delta F_{\text{mean}} \). Equally, considering all the cumulative probabilities \( P(\Delta F_{\text{max}}) \) calculated on the available predictions for each time step (hour) in the training set, they have a standard deviation \( \sigma = 0.025 \), a maximum value of 0.929 and an average value of 0.875. For this reason, in this study each observed \( \Delta F(t) \) having a cumulative probability higher than 0.95 (see Figure 4) has been considered as a possible anomaly.

- The criterion at the previous point for identifying possible anomalies implies that anomalies are necessarily higher than \( \Delta F_{\text{max}} \), but does not consider potential systematic errors in the meter readings. Thus, assuming a certain meter accuracy (e.g., 4% is assumed in this case study), among anomalies only the values of \( \Delta F(t) \) exceeding \( \Delta F_{\text{max}} \) plus that meter accuracy allowance generate alarms, in the sense that the exceedance is beyond a possible error of measurement.

For the sake of brevity, in the following only some of the analysed weeks are discussed extensively. However, all
parameters are reported in the Appendix with reference to Equations (3)–(5). To simplify the visualisation of normal and abnormal situations, all the following figures show the average predicted water consumption $\Delta F^{\text{mean}}(t)$ as a continuous black line, and the values of $\Delta F^{\text{max}}(t)$ (upper line) and $\Delta F^{\text{min}}(t)$ (lower line) as two dotted lines, which model predictions. In Figure 5 (and following figures) the measured $\Delta F(t)$ values are represented by the white dots.

**Figure 4** | Cumulative probability criterion for anomaly identification.

**Figure 5** | 12th week diagram, from the 8th to the 12th (working days) and from the 13th to the 14th (weekend days) of September 2008.
while black circles are possible anomalies and grey squares are possible alarms.

The 12th week – from the 8th to 14th of September 2008

EPR models developed using the first 11 weeks are tested on unseen data in week 12. Figure 5 shows the measured data during the 12th week, thus model (3) has been used to provide 11 predictions of $\Delta F(t)$ (i.e., 11 weeks) for each time step, given 11 different values of $a_1$ and $a_2$. Figure 5 also contains the same diagrams (maximum, minimum, average and measured values) for the two weekend days (Saturday the 13th and Sunday the 14th) of the week considered, calculated using Equation (4). From Figure 5, the 8th of September is characterised by some possible anomalies, which are repeated on the 9th, now with alarms being identified.

The situation seems to deteriorate further during the 11th and 12th with more anomalies and alarms identified, as indicated for example in the zooming window of the 11th consumptions in Figure 5. This suggests a possibility of an anomalous event during these last couple of days. The importance of the event could be proved by the fact that, as indicated in Table 1, there are two customer contacts (and a subsequent intervention) on the 11th and 12th of September.

There are some problems during the 13th, while during the night of the 14th something significant happened, as shown by a series of alarms in the night between the 13th and 14th (see the circle zooming window in Figure 5). This caused a contact from customers on the 14th (see Table 1). During this weekend, there is also a gap in data between 10:00 and 15:00 on the 13th of September due to faulty communication/equipment, but this does not influence the applicability and the efficiency of the procedure.

As an alternative to the proposed approach, it is possible to calculate the average values of water consumption for every time step based on the 11 weeks used for training, and the relative minimum and maximum values observed in the training weeks. Figure 6 shows the comparison for the days when there are complaints by customers in this week, i.e., the 11th, 12th of September (Figure 6, left) and the 14th of September (Figure 6, right).

According to the same criteria adopted for analysis of the EPR models, as reported above, there is only one anomaly and one alarm on the 11th September, which cannot be considered enough to indicate with certainty a malfunctioning in the network. Note that in Figure 6 (left), the observed values for the 11th of September are white circles, while observed values for the 12th of September are white triangles. In Figure 6 (right), the observed values for the 14th of September are white circles.

The 13th week – from the 15th to 21st of September 2008

Among the optimal models returned by EPR for this second testing sub-phase, the selected models have the same structure (number of terms and selected inputs) of Equations (3) and (4) for working days and weekend, respectively, with slight differences in the parameter values (see Tables A1 and A3).
The diagram in Figure 7 shows that going from Monday (15th) to Tuesday (16th) the number of anomalies increases, with a possible alarm (square dot) in the morning. This could indicate that something strange is happening in the network. In the night between the 16th and 17th, in fact, a slight increase in MNF is recorded, and during Wednesday (17th) anomalies and alarms are confirmed. This sequence can justify an intervention on the asset in order to fix the possible problems/inefficiencies. In fact, Table 1 reports a couple of interventions on the 16th and 17th of September (it is, however, not known if these interventions have been planned or executed as repairs of problems that have occurred). Going to Thursday (18th), the problem on MNF has been fixed, while anomalies are still present, even if they reduced on Friday (19th), Saturday (20th) and Sunday (21st) (see Figure 7). What is likely to have happened is that the works on the 16th did not solve the whole problem, thus the MNF slightly increased after the works; then, the works done on 17th fixed most of the problem, since the following MNF returned to the usual values.

The 15th week – from the 29th September to the 5th of October 2008

This week is important since a significant event for the DMA was registered. The structures in Equations (3) and (4) are confirmed for this week, with parameters provided in Tables A1 and A3 (see Appendix, available with the online version of this paper). During this week, possible anomalies (and alarms) are increasing from Monday the 29th of September to Wednesday the 1st of October, with a slight increase of the MNF (see the zooming windows in Figure 8). On the 1st of October, the criticalities appear reduced as well as the MNF between the 1st and the 2nd, which returned to previous values. In fact, Table 1 reports a customer contact during the 1st of October, which likely reduced the entity of the problems of the network. The anomalies grow again from Friday the 3rd to Saturday the 4th of October, with an increase of MNF, and around 9 pm on Sunday the 5th, some high measures of customer consumption are registered as indicated by a sequence of grey square dots on the right of Figure 8. This situation is a strong candidate to indicate that a big event happened in the network, also in

Figure 7 | 13th week diagram, from the 15th to the 19th (working days) and from the 20th to the 21st (weekend days) of September 2008.
relation to the sequence of anomalies in the days before, but needs to be confirmed from further evidences.

The 16th week – from the 6th to 12th of October 2008

Also for this week, among the Pareto front of models returned by EPR, the above reported model structures are still preferred for both working days and weekend, with coefficient reported in Tables A1 and A3.

As noted at the end of Sunday the 5th of October, something happened in the network after 9 pm. The observed values of customer consumption are generally higher than those predicted by the EPR models (i.e., that model the history of the network consumptions), with many alarms (grey squares), indicating that the observed values largely exceed the maximum predicted values (by at least 4%), as shown in Figure 9. This is the evidence needed to state that a great change in water withdrawal is happening, likely due to an unreported burst. The importance of the burst/leak in the network is confirmed by the increased MNF (from 6.5 to 8 L/s).

During this week, there is also customer contact on the 8th (likely due to pressure problems due to the new leak occurring), and some crew interventions: one on the 6th and two on the 9th, through which technicians have tried to limit the problems. Nonetheless, the higher network outflow observed during this week has a masking effect on any other possible (smaller) unusual behaviour of the network, because the training set (the 15 previous weeks) does not contain similar consumptions, thus the ‘history’ of the network returns possible maximum values of water consumption lower than those recorded in this week.

The 17th week – from the 13th to 19th of October 2008

The EPR model search procedure still contains the same reported structures. This is the first test week for which the training set includes a previous week with a large leak (the 16th). It is worth noting that when the training weeks do not contain significant leaks in their records (from the 1st to the 15th week), the average standard deviation $\sigma$ of the EPR predictions for each time step is almost small (0.172 L/s), and thus the lines of maximum and mean EPR predicted values are quite close to each other. This way, possible anomalies and alarms can be identified by focusing the attention on points above the maximum values line, as the adopted criteria aforementioned try to do.
Including weeks with big leaks into the training set causes the increase of the average standard deviation $\sigma$ of the EPR predictions for each time step ($0.333$ L/s for the 17th week), as shown by the increased distance between the average value line and the maximum value line, since the presence of significant leaks can influence the maximum predicted values of $\Delta F(t)$. This is also evident by observing the coefficients $a_1$ and $a_2$ for the model of working days (Table A2 in the Appendix), and the coefficients $a_3$ and $a_4$ for the weekend model (Table A4). Coefficients of the last model, referring to the 16th week, are the highest, thus the EPR model for the 16th week (which contains the large leak) returns the maximum values among the 16 predictions. This evidence should lead to reconsidering the criterion of discrimination of anomalies. In fact, the hitherto used criteria for identification of possible anomalies (i.e., $P(\Delta F(t)) > 0.95$) and alarms (i.e., $\Delta F(t) > 1.04\Delta F_{\text{max}}$) are focused on the maximum value line. According to the aforementioned observations, this means that the inclusion within the training weeks of ‘out of normal’ weeks (with big leaks) raises the bar beyond which a future observation has to go to be considered an anomaly/alarm. In a future perspective, when consumptions hopefully will return to ‘normal’ values, the hitherto used criteria would lead to identifying anomalies only if there are observed values close to maximum predictions (i.e., reproducing historic highs), ignoring a set of values, not being greater than maximum predictions but significantly higher than the average predictions $\Delta F_{\text{mean}}$.

For this reason, the identification criteria for possible anomalies and alarms has been reconsidered and the following assumptions have been made:

- Given the Normal distribution for the EPR predictions, with average $\Delta F_{\text{mean}}$ and a standard deviation $\sigma$, every value of observed $\Delta F(t)$ having a cumulative probability higher than $P(\Delta F_{\text{mean}} + \sigma) = 0.84$ is now considered as a possible anomaly.
- Among possible anomalies, all values of observed $\Delta F(t)$ that exceed $(\Delta F_{\text{mean}} + \sigma)$ over the meter accuracy (4% assumed in this case) are considered as alarms, indicating an abnormal increasing of water withdrawal from the network.

Figure 9 | 16th week diagram, from the 6th to the 10th (working days) and from the 11th to the 12th (weekend days) of October 2008.
This means that the borderline between normal and abnormal conditions become $(\Delta F^{\text{mean}} + \sigma)$, always considering the accuracy of the meter gauge used for alarm raising.

Also for this week (as well as for following weeks 18th, 19th and 20th), all secondary problems are masked by the presence of the unreported burst (there are two interventions on the 13th and 16th and a customer claim on the 16th), since for the most part the observations are close to or exceed the maximum value line (see Figure 10). As will be shown by the analysis of the 21st week, when things go back to normal, the observations return to normal values and it is easier to recognise possible anomalies.

**The 21st week – from the 10th to 16th of November 2008**

After 3 weeks characterised by the presence of the large leak, during this week the utility crew finally located and repaired the burst. The same previously reported structures are still present, with reference to Equation (5).

During this week, there are four interventions, two on the 10th and two on the 15th of November; one on Monday the 10th in which the leak was discovered and repaired (likely before 2 pm), as clear from the diagram in Figure 11, and confirmed by the reduction of the MNF during the following nights. After that date, all the recorded water consumption values are below the average value line, with the only exception of a couple of records (half an hour) in the morning of the 14th of November, where likely a (little) abnormal condition has occurred.

For the weekend, all consumption values came back below the average value line, meaning that the network is operating in ‘normal’ conditions, given that such conditions include the background leakages.

**Assessment of possible water volume lost**

An additional aspect of the problem tackled in this application is to try evaluating the water volume loss during the monitored periods. This could be useful in order to...
understand the nature of possible problems encountered during monitoring of WDN operation. According to the above-mentioned assumptions for identification of anomalies and alarms, a range of possible water volume losses has been determined for each day analysed in the test set (10 weeks):

- the estimation of the minimum possible water volume loss in a day ($V_{\text{min}, \text{day}}$) is calculated as follows:

$$V_{\text{min}, \text{day}} = \sum_{N} [\Delta F(t) - TS] \cdot \Delta t$$

$$\begin{cases} 
TS = \Delta F_{\text{max}} & \text{from 12th to 16th week} \\
TS = \Delta F_{\text{mean}} + \sigma & \text{from 17th to 21st week}
\end{cases}$$

where $N$ is the number of sampling time steps available in a day; $\Delta t$ is the interval between two measures; and $TS$ is the threshold used for identifying anomalies.

- the estimation of the maximum possible water volume loss in a day ($V_{\text{max}, \text{day}}$) is calculated as follows:

$$V_{\text{max}, \text{day}} = \sum_{N} [\Delta F(t) - \Delta F_{\text{mean}}] \cdot \Delta t$$

assuming as pejorative hypothesis that all observations exceeding $\Delta F_{\text{mean}}$ were actually abnormal (i.e., due to new leakages occurring).

Figure 12 reports the trend of these values in the observed period, where estimations for working days and weekends are reported in chronological order. From Figure 12, there is a relevant event on the 14th of September (water volume loss between 15 and 20 m$^3$/day, thus with a leak flow rate between about 0.15 and 0.25 L/s), as also evident by anomalies in Figure 5 and an intervention due to customer contact (Table 1) that has reinstated the normal network operational conditions.

The period between the 25th of September and the 1st of October indicates some problems (as confirmed by recurrent anomalies in Figure 8). In these days, there have been a couple of customer contacts (and a consequent intervention, see Table 1) that confirmed that there was a problem (a water volume loss between 10 and 25 m$^3$/day, thus with a leak flow rate between about 0.10 and 0.25 L/s) that continued after the 2nd of October (see the increasing trend of water volume lost from the 2nd to the 5th of October). What
seemed to be brewing during the ten previous days has hap-
pened during the night between the 5th and the 6th of
October, causing a large water loss (a water volume lost
between 90 and 110 m$^3$/day, thus with an average leak
flow rate between 1 and 1.2 L/s a day). In the following
month, utility crews probably tried to contain and locate
the water loss, as witnessed by many interventions after
the 5th of October, until they identified and repaired
(likely) the large break on the 10th of November (see
Figure 11).

It is worth noting that the leakage evaluation method-
ology here reported cannot consider (nor can it estimate)
previously existing background leakage in the analysed
WDN; this means that null (or quasi-null) values after the
10th of November mean that things were back to the
previous ‘normal’ operation with existing background
leakages.

CONCLUSIONS

This paper investigates the effectiveness of the EPR modell-
ing paradigm to reproduce and predict a WDN
consumption behaviour using on-line measured data by
low-cost pressure/flow devices. The proposed application
shows the ability of the EPR to return physically consistent
formulations for representing the consumption ‘history’ of a
real-life DMA. This ability was presented in the case study,
which includes a limited number of data flow/pressure
measurements collected during normal daily WDN operat-
ing conditions, in a time window of 21 weeks.

The EPR modelling paradigm was used to detect
anomalies due to possible unreported bursts in the network
using pressure/flow measurements at a few points. This
analysis was supported by other data, such as customer con-
tacts due to problems in water service delivery and utility
intervention records. The consistency of possible problems
has been analysed also considering the MNF variation regis-
tered. The procedure allows the estimation of a possible range
of water volume lost, which is an output of the methodology.

The procedure proved to be effective in detecting poss-
ible anomalies and the presence of a large leak, even if it
cannot identify the location of the burst. The procedure,
which requires simple data pre-processing and no subjective
interpretation of results, is suitable to be readily automated
and connected with the data acquisition systems (SCADA)
available to water companies. Furthermore, if coupled
with leak location methodologies, the presented procedure
can contribute to the reduction of false alarms, allowing
smaller intervention times and therefore reducing customer
claims. A drawback is that the presence of a significant leak
can mask the occurrence of smaller leaks until the former is
discovered and repaired. However, the inclusion of weeks
with large leaks among the training set makes the procedure
more robust with respect to future similar unreported bursts.

From a data-driven modelling standpoint, using the EPR
modelling approach has a number of advantages with respect
to other data-driven techniques (e.g., ANNs), such as the
following:

- The model construction and the selection among candi-
date explanatory variables is automatically performed,

Figure 12 | Estimation of possible water volume losses during the period from 8th September to 14th November 2008.
without previous assumptions about the form of the model equation by the user. This also indicates which useful variables can be conveniently observed during future monitoring campaigns, as well as about the appropriate sampling time step.

- The EPR multi-objective paradigm returns a set of models that can be compared in terms of both selected variables (i.e., past time steps) and error statistics, thus avoiding over-fitting to past data. Such models are linear with respect to regression parameters, thus allowing easy analysis of their uncertainty over time.

- The range of predictions obtained by the MCS EPR modelling strategy reflects different customer behaviour over time (i.e., weekly demand/pressure patterns), instead of purely probabilistic assumptions, thus implicitly including all the uncertainty surrounding water demand and background leakages.

The structure of the preferred EPR models are composed of two terms, one related to network inflow rate and one related to average pressure on the network, thus it can be related also to background leakages as stated by several works in the field (e.g., May 1994). Future works could investigate the nature of such components of EPR models, also on other databases, for a possible direct assessment of the water loss starting from field data.

ACKNOWLEDGEMENTS

The research reported in this paper was founded by two projects of the Italian Scientific Research Program of National Interest PRIN-2012: ‘Analysis tools for management of water losses in urban aqueducts’ and ‘Tools and procedures for advanced and sustainable management of water distribution networks’.

REFERENCES


WRc, 1994 Managing Leakage, Report A. Water Research Centre Bookshop, Swindon, UK.


First received 27 May 2015; accepted in revised form 7 November 2015. Available online 12 December 2015