International Conference on Hydroinformatics

2014

Detecting Pipe Bursts In Water Distribution Networks Using EPR Modeling Paradigm

Luigi Berardi

Daniele Laucelli

Orazio Giustolisi

Dragan A. Savić

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

DETECTING PIPE BURSTS IN WATER DISTRIBUTION NETWORKS USING EPR MODELING PARADIGM

LUIGI BERARDI(1), DANIELE LAUCELLI(1), DRAGAN SAVIC(2)

(1): Civil Engineering Department, Technical University of Bari, Via E. Orabona 4, Bari, 70125, Italy

(2): College of Engineering, Mathematics and Physical Sciences, University of Exeter, North Park Road, Exeter EX4 4QF, UK.

Efficient management of water distribution systems requires effective exploitation of available data from pressure and flow devices. This also means that water companies need to balance the progressively increasing amount of information available and actually usable with the cost of gathering data. Among different techniques developed in the last few decades, those implementing data mining for analyzing pressure/flow data appear very promising. This is because they rely on empirical observations of the system behavior over time, without detailed knowledge of pipe network flows and pressures. This paper investigates the effectiveness of the evolutionary polynomial regression (EPR) paradigm to reproduce system behavior using on-line measured data by cheap pressure/flow devices. Using data from a real district metering area, the present case study shows that EPR can be effective in reproducing the behavior of the water system from available flow/pressure measurements. The output can then be used for various purposes and, in particular, to detect anomalies due to possible unreported bursts. Such an EPR model might be integrated into an early warning system to raise alarms when anomalies are detected.

INTRODUCTION

Sustainable management of water distribution networks (WDNs) requires the timely detection of water leakages from pipelines. This can reduce waste of a precious resource, decrease cost of treatment and pumping, cut third party damage and, ultimately, reduce greenhouse gas emissions. To this end, timely detection and location of pipe bursts in a WDN is really important. Pipe bursts represent a potential risk to public health and can cause significant environmental damage and economic loss. Despite all the advancements made in methodologies for burst detection and location, further improvements of their efficiency and reliability are still needed [1].

Currently, solutions to burst detection and location problems employ highly specialized hardware equipment, such as leak-noise correlators [2] and pig-mounted acoustic sensing [3], which are the most accurate in today's bursts detection and location surveys [1]. However, they can be expensive and time demanding, or even require the shutdown of pipeline operations for long time periods. As a consequence, water utilities are asking for faster, cheaper and manageable techniques. Water companies commonly utilize hydraulic sensor technology and on-line data acquisition systems, which enable them to deploy a large number of accurate and cost effective pressure and flow devices. The data collected by these devices provide a potentially useful source of information for quick and economic detection and location of pipe bursts in WDNs.

Therefore, a number of numerical techniques that attempt to efficiently employ these data have been recently developed (e.g., Liggett and Chen [4]). Among the numerical techniques, those using data mining and other AI tools appear very promising for automatic on-line analysis of the pressure and/or flow data (e.g., Mounce *et al.* [5]; Romano *et al.* [6]). This is mainly because such techniques rely on empirical observations of the WDN behavior over time, without the need for detailed knowledge of the pipe network (e.g., through hydraulic modeling or asset parameters). These procedures are based on: (i) data preparation (e.g., de-noising; data reconstruction); (ii) prediction of expected values based on data-driven models; (iii) identification of anomalies in flow/pressure and raising alerts based on a mismatch between model predictions and signals from meters.

Among available AI tools, the present paper analyzes the potential of the evolutionary polynomial regression (EPR) modeling paradigm in this framework. The idea is to use the Multi-Case EPR Strategy [9][10] to develop a water consumption prediction model using values recorded over a number of past time windows (i.e., weeks) that are treated as separate data-sets. This means to develop the same mathematical structure (i.e., formula) shared by several prediction models, but with different sets of parameters, each minimizing the error over a different time window. As it will be made clear in the following sections, this results in a range of predictions for the system water consumption given the pressure/flow measurements in a few points of the WDN. These predictions can then be used to detect anomalies and for raising alarms. The methodology is tested on data coming from an engineered experiment on a real district metering area (DMA).

MODELING APPROACH: MULTI-CASE EPR

Evolutionary polynomial regression (EPR) is a hybrid modeling technique that allows the exploration of polynomial models, where candidate covariates are included in the final model based on the accuracy of predictions and parsimony of the symbolic model expression [7]. A pseudo-polynomial structure for model expression is used, where each term comprises a combination of candidate inputs (covariates); each covariate 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 minimizing the error on training data. Each monomial term can include user-selected functions among a set of possible alternatives.

The search problem is defined in a multi-objective optimization framework, where candidate models are evaluated based on three criteria, namely, (a) model accuracy (maximization of fitness to data), (b) parsimony of covariates (minimizing the number of explanatory variables included in final model expressions) and (c) parsimony of mathematical equation (minimization of the number of polynomial terms). While the accuracy criterion for model evaluation is intuitively understood, the role of the parsimony criteria in EPR aims to prevent over-fitting of model to data, and thus endeavor to capture underlying general phenomena without replicating noise in data. In this way, the technique allows the most important input covariates for the phenomena under study to be identified. The EPR uses a multi-objective genetic algorithm (MOGA) optimizer to find candidate models and rank them utilizing the above mentioned criteria and the Pareto dominance methodology (Giustolisi and Savic [7]).

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., a covariate raised to the power of zero is *de facto* excluded from the model [8]. 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, 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 analyzed. In addition, the recurrent presence of certain covariates in several non-dominated models indicates the robustness of these covariates as potential explanatory variables of the phenomenon. All these features make the EPR modelling paradigm substantially different from purely regressive methods (e.g., artificial neural networks, ANN) 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 [9].

When the available data refer to different realizations of a certain physical phenomenon under various conditions/observations, it can be more difficult to identify the pattern among variables describing the underlying (i.e., main) system behavior [10]. The Multi-Case EPR is suitable for situations where data can be partitioned into subsets, each representing a particular realization/experiment of the same phenomenon. Thus, the Multi-Case EPR simultaneously identifies the best pattern among significant explanatory variables describing the same phenomenon in all data partitions, while neutralizing possible impacts of errors and uncertainty in data. The Multi-Case EPR also makes use of the MOGA optimization scheme, as described above, where each candidate model structure is evaluated on each considered data partition [9].

METHODOLOGY APPLICATION

In this study, the Multi-Case EPR is applied to develop a water consumption prediction model using pressure/flow measurements recorded over a period of time in an urban DMA in the UK. The available database can be considered as a time series, with a measurement time step of15 minutes, thus a full day observation consists of 96 records. Among the available data some selection and pre-processing has been performed, as detailed in the following sub-section. The used data has been divided into a number of weekly datasets, assuming that the observed phenomenon (i.e., water consumption) has different realizations on a weekly basis.

The goal of the exercise is to predict the presence of anomalies in the DMA behavior (i.e., possible unreported bursts) by means of a water consumption model built on pressure/flow measurements at a few points in the DMA. It is expected that the use of Multi-Case EPR on weekly data will lead to a common mathematical structure for the prediction model as being representative of the underlying phenomenon for all weekly datasets.

Thus, every dataset has different model coefficients and the same model structure; this leads to a range of predictions for the DMA water consumption, one for each analyzed weekly dataset. The range of predictions reflects different past time behavior of customers (i.e., weekly demand/pressure patterns). This is an alternative approach to purely probabilistic models that are usually implemented to consider uncertainties in water consumption, the presence of background leakage and possible measurement errors. If observed values of water consumption are outside the range predicted by the model, the system is assumed to be experiencing an anomaly, as it deviates from the expected behavior. This could be due to abnormal water usage or unreported bursts. In the present case, the abnormal functioning is caused by an engineered event, which is performed to reproduce a pipe burst in the network, consisting of a hydrant opened for 24 hours with a constant water outflow (2 l/s), see Table 1.

Data collection and pre-processing

The available input data consists of a time series of pressure (meters) and flow values (liters/second), measured at a time step of 15 minutes at the inlet measurement point (P_{31} and F_{31}, respectively), at the outlet measurement point (P_{32} and F_{32}, respectively) and at the internal measurement point (P_{33}) of a DMA. The output data (ΔF) consists of a time series of water consumption (liters/second) calculated as the difference between the water flow at the inlet point and the water flow at the outlet point of the DMA; thus, the water consumption is here accounted for as an average flow over the considered time step, instead of water volume. Figure 1 shows the DMA layout, the location of measurement points, the engineered event hydrant and the elevation of nodes in the analyzed DMA.

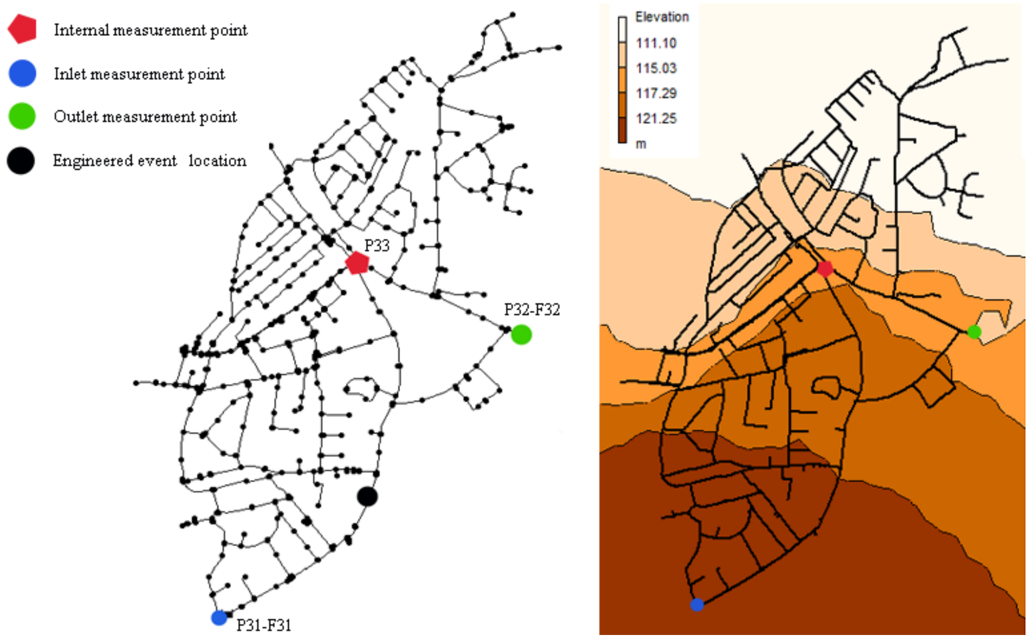

Figure 1. Layout and elevations of the analyzed DMA.

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.

The time windows of the available data are: from 16 June 2008 to 31 August 2008; from 8 September 2008 to 5 April 2009; and from 18 May 2009 to 21 June 2009, for a total amount of 55 weeks. Every weekly dataset has 672 data points (the time step is 15 minutes). Some weeks have been omitted due to a number of gaps in data records, thus the total number of usable weekly dataset is 41 (from Monday to Sunday).

After the observation period, the water utility simulated a number of engineered experiments by opening fire hydrants for 24 hours in different locations in the DMA, thus causing a change in the hydraulic state of the system. During the engineered events, pressures and flows at the measurement points were recorded using the same time step. In the present work only one event is considered, whose main features are summarized in Table 1.

Table 1. Details of the used engineered event in the analyzed DMA.

Event	Date	Size (l/s)	Hydrant Opened at	Hydrant Closed at
C	24th July 2009	2	17:30 24/07/2009	16.00 25/07/2009

Results and discussion

For a Multi-case EPR run the candidate values chosen for the exponents were [-2,-1.5,-1,-0.5, 0, 0.5, 1, 1.5, 2], thus exploring well known possible relationships, e.g., linear, quadratic, inverse linear, square root, etc., for each term involved in the models. Such a modelling choice was mainly aimed towards finding more general formulations. The number of past time steps considered for input data was set to 4, while no past values of output data were used. The number of polynomial terms was set to m = 4, plus the bias (i.e., a constant value in the polynomial expression). The MOGA process was set to run for 5,000 generations. The objective functions optimized are: (1) maximization of model accuracy as measured by the means of the coefficient of determination (CoD) [8]; (2) minimization of the number of explanatory variables; and (3) minimization of the number of polynomial terms.

The EPR returned 17 models, all optimal in a Pareto dominance sense. Almost all the models have a very high accuracy to training data (i.e., average CoD = 0.924) and contain recursively two out of five explanatory variables: the pressure and flow measured at the inlet measurement point (P_{31} and F_{31}). This indicates the importance of the inlet measurement point for modelling the system behavior. Thus, in order to adopt a simple expression, as a trade-off between accuracy and parsimony, the following model structure has been chosen:

$$\Delta F(t) = a_1 F_{31}(t) + a_2 \sqrt{P_{31}(t-1)} \cdot F_{31}(t) + a_3 \sqrt{P_{31}(t)} + a_0 \qquad (1)$$

where, P_{31} $(t-1)$ is the pressure head at the inlet measurement point at 1 time step before the time t. The selected model achieved an average CoD of 0.994, whereas for each single dataset a string of 4 coefficients (a_3, a_2, a_1, a_0) is determined. This allows the calculation of 41 water consumption predictions $\Delta F(t)$, one for each of the models associated with the structure in Eq. (1). These predictions are representative of the water consumption history of the system; this means that, given the observed values of P_{31} and F_{31}, the model (1) is able to predict possible values of water consumption, according to past behavior. From this perspective, the value ΔF^{mean} calculated as an average of the 41 predictions, can be considered as the most probable expected water consumption at time t given the two inputs P_{31} and F_{31}, according to Eq.(1). If the observed value of $\Delta F(t)$ is close to ΔF^{mean} this means that it is consistent with the history of the system, and can be considered as a normal functioning condition.

Similarly, if the observed value of $\Delta F(t)$ is less than ΔF^{mean} this means that at time t customers are withdrawing less water than usual, and this does not result in an alarm; conversely, when observed value of $\Delta F(t)$ is above ΔF^{mean}, customers are withdrawing more water than expected, with respect to past history of the system; this condition becomes more critical when the observed value of water consumption is close to (or above) the predicted maximum value, ΔF^{max}, thus leading to a suspected anomaly (i.e., unreported burst) in the network. From a statistical point of view, an observed value of water consumption that is on the higher limit of the range of predictions (i.e., equal to $\Delta F^{\max}$) has the probability of only 1/41 for the system to be in an acceptable functioning condition.

To ease this comparison, Figures 2 and 3 show a continuous black line that represents the average predicted water consumption $\Delta F^{\mathrm{mean}}(t)$, and two dotted lines representing the value $\Delta F^{\max}(t)$ (upper line) and the value $\Delta F^{\min}(t)$ (lower line) as calculated by the predictions of the model.

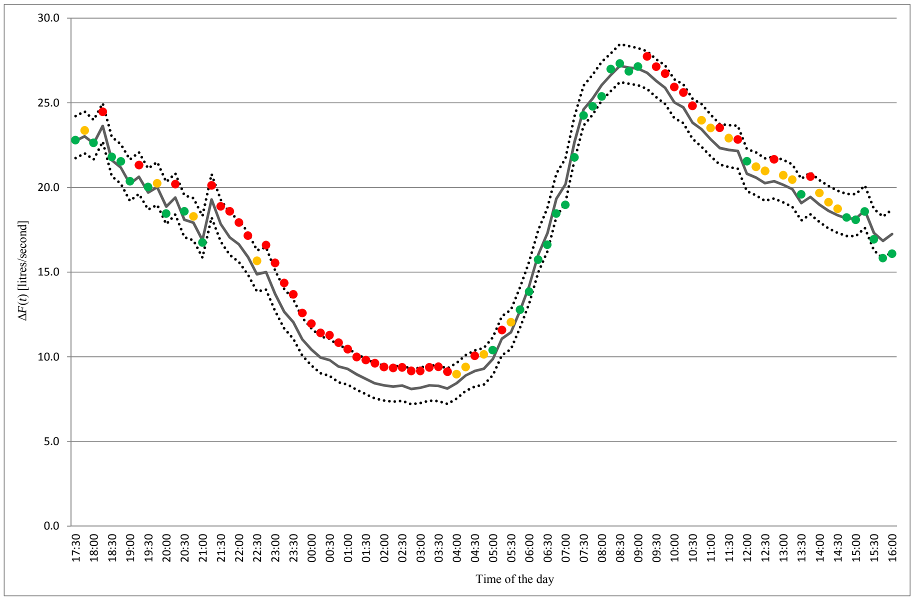

Figure 2. Diagram of $\Delta F(t)$ for the event C.

Figure 2 represents the range of EPR model predictions for the engineered event C and the measured values of ΔF during the event. Note that the prediction range in Figure 2 indicates that, given the observed values of P_{31} and F_{31} during Event C, the most expected values of $\Delta F(t)$ are shown as the points on the black line, according to what was experienced by the system in the past; being below that line indicates conditions less critical that expected, while points above the black line indicate possible alarms. In Figure 2, the green dots indicate observed values of water consumption that can be considered normal for the analyzed DMA; the yellow dots indicate conditions above the average; and the red dots indicate possible alarms. It is clear from Figure 2 that most potential alarms occur during night-time, when water demand is the lowest and the values of water consumptions in the past are basically constant. However, in case of event C, there are also many potential alarms just after the morning consumption peak.

The left hand graph in Figure 3 reports the ΔF values for a previous day of Event C (the middle graph, from 17:30 to 16:00) and values (the right hand graph) for the following day in July. In these 2 days, when no recognized loss event has occurred, all the observed values of water consumption are located between the average value ΔF^{mean}, and the minimum value $\Delta F^{\min}$; this may indicate that customers are behaving in an average manner, based on the previous observed weekly patterns.

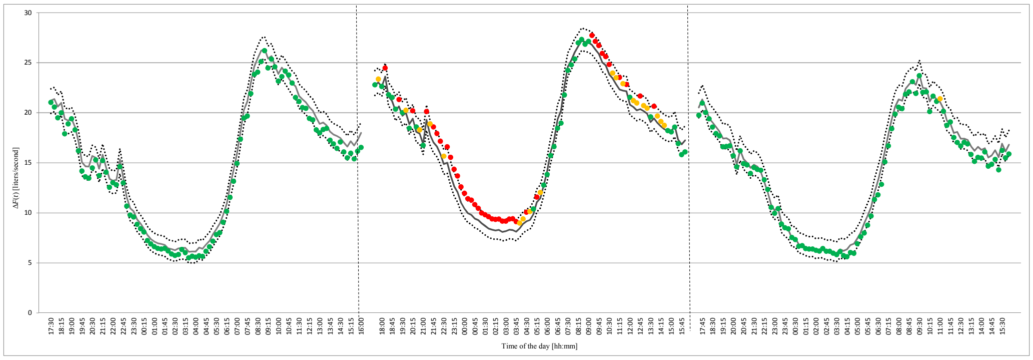

Figure 3. Diagram of $\Delta F(t)$ for event C compared to one previous day (left) and one following day (right).

Generally speaking, even when slight differences due to the different period of the year (the first available previous day was in May) are ignored, the two ordinary days present very similar trends in water consumption, especially during nighttime (about 5 l/s). Conversely, during event C the diagram appears to be shifted upwards, as it is clear during the peak hours. In particular, during nighttime the predicted values of ΔF are greater than the average by at least 2 l/s, considering the lower curve, which is the value of the engineered event (leak) at the hydrant.

This behavior could be expected since the model is able to reproduce the response of the system given the pressure/flow at the inlet. The key point is that, having a set of 41 parameters (i.e., models) it possible to calculate a range of predictions for each time t, representing the past patterns, where the average value of such predictions represent the most probable expected value. Being above or below this value means that the observed values “have happened” more rarely in the past, since the coefficients represent all the uncertainty about water demand, leakage and measurement errors.

CONCLUSIONS

The paper presents an initial investigation into the effectiveness of the EPR modelling strategy to reproduce the WDN behavior using on-line measured data made available by cheap pressure/flow devices. The proposed application shows a promising ability of the EPR model to perform unreported burst detection in a real-life DMA, even considering a limited number of data measurements.

In the reported case study EPR showed the following strength points with respect to other data-driven techniques (e.g., artificial neural networks):

(i) the model construction and the selection among candidate explanatory variables is automatically performed by EPR, without previous assumptions by the user;

(ii) the number of past time steps to be used for prediction is selected automatically by EPR from the available set; these point (i) and (ii) result in indications about the variables that can be conveniently observed during future monitoring campaigns, as well as about the appropriate sampling time step;

(iii) 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;

(iv) the returned models are essentially linear with respect to regression parameters, this allows easy analysis of their uncertainty over time;

(v) the range of predictions obtained by the Multi-Case EPR modelling strategy reflects different customer behavior 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.

Acknowledgments

Authors wish to acknowledge Dr. Michele Romano for his helpful suggestions and guide for the data pre-processing phase of this work.

REFERENCES

[1] Puust R., Kapelan Z., Savić D.A. and Koppel T., "A review of methods for leakage management in pipe networks", *Urban Water Journal*, Vol. 7, No. 1, (2010), pp 25-45.

[2] Grumwell D., and Ratcliffe B., "Location of underground leaks using the leak noise correlator," *Water Research Centre*, Technical Report 157, (1981).

[3] Mergelas B. and Henrich G., "Leak locating method for pre-commissioned transmission pipelines: North American case studies", *Proc. Leakage 2005*, Halifax, Canada, (2005).

[4] Liggett J.A. and Chen L., "Inverse Transient Analysis in Pipe Networks," *Journal of Hydraulic Engineering*, Vol. 120, No. 8, (1994), pp 934-955.

[5] Mounce S.R., Boxall J.B. and Machell J., "Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows", *Journal of Water Resources Planning and Management*, Vol. 136, No. 3, (2010), pp 309-318.

[6] Romano M., Kapelan Z. and Savić D.A., "Real-time leak detection in water distribution systems", *Proc. WDSA 2010*, Tucson, Arizona, (2010).

[7] Giustolisi O. and Savic D.A., "Advances in data-driven analyses and modelling using EPR-MOGA", *Journal of Hydroinformatics*, Vol. 11, No. 3, (2006), pp. 225-236.

[8] Giustolisi O. and Savic D.A., "A symbolic data-driven technique based on evolutionary polynomial regression", *Journal of Hydroinformatics*, Vol. 8, No. 3, (2006), pp. 207-222.

[9] Savic D.A., Giustolisi O. and Laucelli D., "Asset performance analysis using multi-utility data and multi-objective data mining", *Journal of Hydroinformatics*, Vol. 11, No. 3-4, (2009), pp 211-224.

[10] Berardi L. and Kapelan Z., "Multi-Case EPR strategy for the development of sewer failure performance indicators", *Proceedings of the World Environmental & Water Resources Congress*, Tampa Bay, USA, (2007) pp 1-12.