Pattern matching and associative artificial neural networks for water distribution system time series data analysis

S. R. Mounce, R. B. Mounce, T. Jackson, J. Austin and J. B. Boxall

ABSTRACT

Water distribution systems, and other infrastructures, are increasingly being pervaded by sensing technologies, collecting a growing volume of data aimed at supporting operational and investment decisions. These sensors monitor system characteristics, i.e. flows, pressures and water quality, such as in pipes. This paper presents the application of pattern matching techniques and binary associative neural networks for novelty detection in such data. A protocol for applying pattern matching to automatically recognise specific waveforms in time series based on their shapes is described together with a system called Advanced Uncertain Reasoning Architecture (AURA) Alert for autonomous determination of novelty. AURA is a class of binary neural network that has a number of advantages over standard artificial neural network techniques for condition monitoring including a sound theoretical basis to determine the bounds of the system operation. Results from application to several case studies are provided including both hydraulic and water quality data. In the case of pattern matching, the results demonstrated some transferability of burst patterns across District Metered Areas; however limitations in performance and difficulties with assembling pattern libraries were found. Results for the AURA system demonstrate the potential for robust event detection across multiple parameters providing valuable information for diagnosis; one example also demonstrates the potential for detection of precursor information, vital for proactive management.

Key words | asset monitoring, auto-associative neural network, event detection system, pattern matching, water distribution systems

S. R. Mounce (corresponding author)
R. B. Mounce
J. B. Boxall
Pennine Water Group,
Department of Civil and Structural Engineering,
University of Sheffield,
Sheffield,
S1 3JD,
UK
E-mail: *s.r.mounce@sheffield.ac.uk*

T. Jackson
J. Austin
Advanced Computer Architecture Group,
Department of Computer Science,
University of York, Deramore Lane,
York,
YO10 5GH,
UK

INTRODUCTION

Population growth, urbanisation, industrialisation and climate change are placing increasing pressure on water resources. The water-energy-food nexus is a term being used to describe the complex linkages and dependencies among water, energy and food security (Olsson 2012). Global demand for water is forecast to outstrip supply by 40% by 2030 due to factors such as population growth and climate change (Parliamentary Office of Science and Technology 2011). This building pressure on water availability is driving a greater consideration of optimal management of clean water resources. Continuous online monitors and sensors are increasingly being used to measure a wide range of potable water hydraulic and quality variables within water distribution systems (WDSs) (Wu *et al.* 2011). Obtaining system information from these data can facilitate proactive system operation and maintenance. For water quality in particular, online data are generally not as reliable as laboratory-based discrete sample analysis with many associated problems that include absolute accuracy, maintenance, calibration, connectivity issues and local disturbances (Aisopou *et al.* 2012). This situation is compounded by the ever increasing volumes of data being collected at a higher than ever seen before frequency of sampling and with coverage of hundreds or even thousands of sites.

doi: 10.2166/hydro.2013.057

Data from online monitors potentially provide a wealth of information about what is happening within WDSs and intelligent algorithms can be applied to turn these data into information for water utility companies. Many companies are not making effective use of what is being collected in this regard and are missing an opportunity to better understand and assess current system status. Any data interpretation system employed must be able to deal with 'dirty data' such as inherent, though improving data variability and quality limitations. Hence systems need to include strategies for handling missing values and dealing with noise, e.g. Branisavljević *et al.* (2010). Analysis systems need to provide useful classifications of system status, events and conditions and not provide an onerous amount of alerts or alarms to system operators who will otherwise ignore warnings hence compromising the value of the information.

This paper presents the application of pattern matching techniques and binary associative artificial neural networks (ANNs) for novelty detection in time series data collected from WDSs. Algorithms are described and a protocol developed for applying the approach to case study data, both hydraulic and water quality, from water supply systems in the UK.

APPROACHES FOR EVENT DETECTION IN WDS MEASURED TIME SERIES DATA

A water distribution network is a complex, distributed, non-linear dynamic system, and thus it may not be effectively or satisfactorily described using purely linear methods or models. It is not possible to build an accurate non-linear model completely describing the system from data due to the uncertainties present. However, data-driven modelling is highly applicable. It has the advantage of not requiring a detailed understanding of the interacting physical, chemical and/or biological processes that affect a system before model inputs can be mapped to outputs. Data-driven models can complement and sometimes replace deterministic models (Solomatine 2002). Recent developments in the field of computational intelligence (sometimes termed soft computing or machine learning) are helping to solve various problems in the water resources domain.

A number of approaches from the fields of artificial intelligence and statistics have been applied for detecting abnormality in WDSs from time series data. Alert systems that convert flow and pressure sensor data into usable information in the form of timely alerts (event detection systems) have been developed with a focus on burst detection to help with the issue of leakage reduction. Some of the most recent approaches are summarised in Table 1. Most of these systems are for detecting leaks/bursts at District Metered Area (DMA) level. DMAs are designed to be hydraulically isolated areas that are generally permanent in the system.

Interest is growing in applying similar event detection systems to online water quality measurements, including from WDSs. The detection of anomalous events is of interest for both daily operational management, with a focus on maintaining high water quality, as well as for identification of intentional or 'natural' contamination events. Jarrett *et al.* (2006) explore data processing and anomaly detection techniques for data from WDSs including control charting, time series analysis, Kriging techniques and Kalman filter techniques. They concluded that no single methodology could be judged to always be the best choice. Open source software known as CANARY (McKenna *et al.* 2007) has been developed by the United

Table 1 | Leak event detection techniques applied to DMA data

Technique	Reference
Time Delay Neural Network	Mounce & Machell (2006)
Belief Rule Based System	Xu *et al.* (2007)
Self Organising Map Neural Network	Akselaa *et al.* (2009)
Mixture Density Neural Network and Fuzzy Inference System	Mounce *et al.* (2010)
Kalman Filtering	Ye & Fenner (2010)
Support Vector Regression with Novelty Detection	Mounce *et al.* (2011)
Multilayer Perceptron, Bayesian System and Statistical Process Control	Romano *et al.* (2014)
Principal Component Analysis	Palau *et al.* (2012)

States Environmental Protection Agency for the analysis of water quality time series data. CANARY uses statistical and mathematical algorithms to identify the onset of periods of anomalous water quality data, while at the same time, limiting the number of false alarms that occur. A two-step process is adopted: state estimation for future water quality value and a second stage of residual classification for determination of expected or anomalous value (an outlier).

The aforementioned event detection systems generally have the following features in common:

(i) They learn from training data in some way to make a prediction about expected future values.
(ii) They have some type of methodology or rules for deciding when sufficient deviation from normality constitutes an abnormal event.

Common difficulties with their application include an often large number of parameters to be tuned, poor quality data and how to define the appropriate training data ('normal' data). Failure signatures often overlap with complex spatio-temporal processes that occur in water distribution networks, for example network configuration changes and abnormal demands (such as industrial processes). This makes differentiation difficult. Another limitation of these techniques is that they generally focus only on anomaly detection being interpreted as outlier detection (Hodge & Austin 2004). However, this simplification produces methods that cannot necessarily discover novel patterns formed by subtle changes across multiple variables over multiple time instances. It is hypothesised that precursor features of a smaller magnitude than such outlier thresholds may be present in some sensor time series datasets, which could be potentially picked up before a major failure event (such as a catastrophic burst). Two approaches are considered here for dealing with these difficulties. (i) Pattern matching – i.e. how to identify generalised features of a pattern corresponding to classes of WDS events. (ii) Associative memories – how a monitoring system stores representation of normal distribution system operation and issues warnings when parameters are deviating from this behaviour so as to detect abnormality and possibly precursors.

THEORETICAL BACKGROUND

Pattern matching

The problem of finding patterns of interest in time series databases (termed 'query by content', i.e. to search for an occurrence of a particular pattern within a longer sequence) is an important one, with applications in many diverse fields of science. Application areas include: patterns associated with growth in stock and share prices (Zhang *et al.* 2010), in neuroscience for analysing the nervous system (Fletcher *et al.* 2008), for space shuttle sensor monitoring (Keogh & Smyth 1997) and in transportation for signal timing in traffic management (Mounce *et al.* 2013). In diagnosis and fault detection applications an engineer may wish to query a pattern database in real-time to determine what past situations (contexts) are most similar to the current sensor profile. Pattern matching can thus be used for identifying anomalies in an online monitoring system. As well as detecting that data are abnormal, it is also useful to be able to determine in what way the data are abnormal and ideally to be able to classify the event type which the data correspond to.

One approach (called *sequential scanning* or *subsequence matching*) is to use brute force and 'slide' the shorter query sequence Q against the longer reference sequence R, calculating the error term at each point based on some similarity measure. A number of steps are required for a general scheme in which we consider a univariate signal uniformly sampled in time, which is the case with WDS time series data:

1. *Data pre-processing*. The data must be processed into such a form so that data from different sensors can be compared on a like-for-like basis. The data for each variable may need normalising both with respect to a mean and with respect to the amplitude.
2. *Populating the libraries*. Each library needs to be populated with data from profiles corresponding to that event type. Firstly, the key variables for this event type need to be identified and then profiles for these variables from past events placed into the library. Figure 1 illustrates some example burst profiles from a library for flow and pressure variables (top and bottom, respectively).

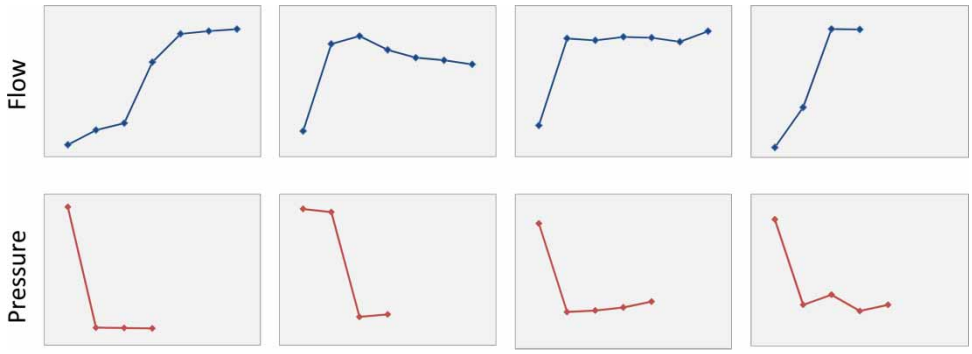

Figure 1 | Burst patterns for WDS hydraulic data (flow top, pressure bottom).

3. *Searching the libraries*. This search is over each of the variables in the data stream. A similarity-based search is used so that similar-shaped profiles of different amplitudes are matched (within a given scaling factor). Matches that are over a given threshold score are returned.

Data pre-processing

It is necessary to pre-process the data in order to be able to compare patterns from different sensors and at different times on a more equal basis. Suppose that the time series for a particular variable is represented by $(x_t)_{t\in T}$. Firstly, we transform the time series to differences from the mean, i.e.

$$x_t \rightarrow x_t - \mu_t \tag{1}$$

for each time t, where $\mu(t)$ is either:

(i) the current average on some moving window $[t-t_A, t]$, i.e.

$$\mu_t = \frac{\sum_{t-t_A\le u\le t} x_u}{n} \tag{2}$$

where t_A is the length of the time window for averaging and n is the number of time series values in the interval $[t-t_A, t]$.

(ii) the average for that time of day, i.e.

$$\mu_t = \frac{\sum_{\substack{t-t_A\le u\le t \\ \tau(u)=\tau(t)}} x_u}{n} \tag{3}$$

where $\tau(u) = \tau(t)$ means that times u and t are at the same time of day and n is the number of measurements u that meet the criteria in the summation.

Secondly, we need to normalise these differences by the standard deviation over the same time series window as the mean is calculated from, so that overall

$$x_t \rightarrow \frac{x_t - \mu_t}{\sigma_t} \tag{4}$$

where σ_t is the standard deviation of the values on which the mean is calculated.

Populating the libraries

The libraries need to be populated with profiles for the different relevant variables. These profiles consist of consecutive measurements over possibly different event window lengths. It is important to use profiles that are typical and indicative of the given event type. A level of expert knowledge and/or water network records may be required to obtain these exemplars.

Searching the libraries

Define $t(E)$ to be the duration of event E. At each time t, the time series profiles used for comparison with event library L are the time series

$$\left\{(x_u)_{t-t_l\le u\le t} | t_l \in \{t(E)|E \in L\}\right\} \tag{5}$$

so that if the event library has profiles of length 30, 60 and 90 minutes, at each time step we would perform a search over the last 30, 60 and 90 minutes worth of data respectively.

The distance between profiles (which must be of the same length in terms of time) is found using the l^2 norm (Euclidean distance), i.e. the distance between the two n-vectors $x = (x_1, x_2, \ldots x_n)$ and $y = (y_1, y_2, \ldots y_n)$ then

$$d(x-y) = ||x-y|| = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{1/2} \quad (6)$$

Match scores can then be generated by calculating (assuming we are comparing profile x with library pattern y),

$$\text{score}(x, y) = \frac{||x|| - d(x, y)}{||x||} \quad (7)$$

so that if $d(x, y) = 0$ then the match score is 1. A threshold can then be used above which two time sequences are said to be similar.

In order to calculate similarity with profiles that are of a similar shape but different magnitude we can calculate instead

$$d_S(x, y) = \min_{\alpha \in [A,B]} d(x, \alpha y) \quad (8)$$

where $[A, B]$ determines the amplitudes of the patterns to compare with, i.e. if this were [0.5, 2] we would compare with patterns of amplitude between half and double that of the library pattern. For practical purposes, the minimum in Equation (8) has to be estimated from a number of intermediate points between A and B.

This research uses this type of pattern matching for populating a pattern library and then comparing a new data stream against it for detecting faults.

Associative memories

Novelty detection is the identification of new or unknown data that a machine learning system has not been trained on or previously seen. Many applications exist for analysing temporal sequences (Keogh *et al.* 2002). Rather than relying on manual review, it is useful to have some form of automated analysis system, which can scan the time series generated by monitoring sensors, and report any abnormal observations. This can be crucial in safety-critical environments. Novelty detection is a two class problem in that it needs to be ascertained whether acquired data come from a normal operating condition or not. There are many techniques for novelty detection including using outlier analysis, however some types of faults do not involve any one variable departing from normal operating range.

Since the classification of novelty is *a priori* unknown, this is a challenging problem and rules out the use of many supervised techniques. There is often no clear-cut boundary between novel events and normal events in real-world applications and a lack of meta-data (such as information about water treatment or process changes, maintenance events, industrial processes, etc.) in WDSs is a particular problem. We can treat the WDS, or sub-areas (such as individual DMAs) in the context of real-time condition monitoring (CM), where it is critical to identify deviations from normal behaviour in sensor readings. A key element of CM is the early detection of potential faults in the monitored system or asset (such as a building, an engine or a pipeline), allowing preventative action to be taken before major damage occurs (for example a catastrophic burst). The CM system has to identify these potential faults based on the values of a (possibly large) number of variables.

In the field of ANNs, an associative memory is a network which stores mappings from specific input representations to specific output representations. Hence, a system that 'associates' two patterns is one that, when presented with only one of these patterns later, can reliably recall the other. There are two types of associative memory: auto-associative and hetero-associative. Auto-associative memories are capable of retrieving a piece of data upon presentation of only partial information from that piece of data, while hetero-associative memories can recall an associated piece of data from one category upon presentation of data from another category. Auto-associative mapping can be created by training an ANN to reproduce its input at its output (Masters 1993). A set of reference signal patterns (e.g. parts of a time series) are learned by the auto-associative network. When presented with an input pattern resembling one of the reference patterns, but

contaminated with noise, the network's output should be close to the learned pattern that most closely represents the trial input. Thus if the trial input were very close to a reference signal (e.g. part of a diurnal flow pattern), with the addition of noise (or missing parts of the signal), the auto-associative network will act as a noise filter (or perform pattern completion). A standard Multilayer Perceptron with D inputs, D outputs and M hidden units can be used in this manner albeit if $M < D$ then additional hidden layers are required to perform non-linear dimensionality reduction (Bishop 2007).

Recurrent networks allow recurrences through feedback connections. This feature is used in associative memories such as the Bidirectional Associative Memory and Hopfield network. The former are ANNs that are used for performing hetero-associative recall (Kosko 1988). Hopfield networks (Hopfield 1982) have been shown to act as associative memories – after a network trains on a set of examples, a new stimulus will cause it to settle into an activation pattern corresponding to the example in the training set that most closely resembles the new stimulus.

This research uses binary associative neural networks for detecting faults, by storing a representation of normal behaviour and monitoring when the asset's activity deviates from this behaviour. They are an example of a hetero-associative memory (although can also be used in an auto-associative fashion).

METHODOLOGY AND SOFTWARE

Signal Data Explorer

The Signal Data Explorer (SDE) is a general purpose data browser and search engine for time series signal data (Fletcher *et al.* 2008). The SDE allows a user to specify the signal event to be searched by supplying a short example of that event (query by content). This can be specified using manually created examples, historical sample inputs, or examples imported from other systems. The user is then able to select the (possibly large) datasets for the search. The search returns a number of potential hits for particular classes of events that can then be browsed using a powerful viewer which assists in the visualisation of multiple signal data channels and enables the user to examine the details of one or more features extracted from the time series data. The SDE also contains efficient spike detection and dynamic filtering functionality. An example of the SDE opening a WDS water quality sensor file is shown in Figure 2. Using the SDE, a user can explore and view any portion of the data rapidly. The output of the data processing tools can be viewed immediately and compared to the raw data by displaying them in the same window. Pattern matching is one of the primary functions of the SDE. The pattern matching functions allow a user to search for particular patterns within or across variables in datasets. The SDE generates a search index based on binary vectors, in a similar fashion to a conventional text search engine.

The SDE provides an interactive and intuitive search capability, which is feature driven, in that the user can highlight a region of interest in a time series signal or select an instance from the pattern template library (as described earlier) and request a pattern matching process to be carried out against the target datasets. Similarity measures are used to provide a ranking system that can score results for the search process. The search process is scalable to terabyte datasets. Application domains for the SDE have included engine vibration data (frequency-power spectra from aircraft engines searching for events such as bird strikes), structural data (e.g. environmental 'shake and bake' tests) and medical data (for finding events in electrocardiography and electroencephalography) (Fletcher *et al.* 2008).

Correlation Matrix Memories (CMMs) and Advanced Uncertain Reasoning Architecture (AURA) Alert

AURA is a set of general-purpose methods for searching large unstructured datasets (Austin 1995). AURA is a class of binary neural network built on CMMs, as an underpinning technology for efficient, scalable pattern recognition in complex and large scale CM applications. During asset operation, the current state of the system can be compared to the stored normal operating behaviour (in the CMM) to see if that combination of variable values has been seen previously. If not, this could be indicative of a problem, even if no individual variables have deviated from their normal value range (Austin *et al.* 2010). Firstly, a quantisation process (binning) is used with each potential value for each

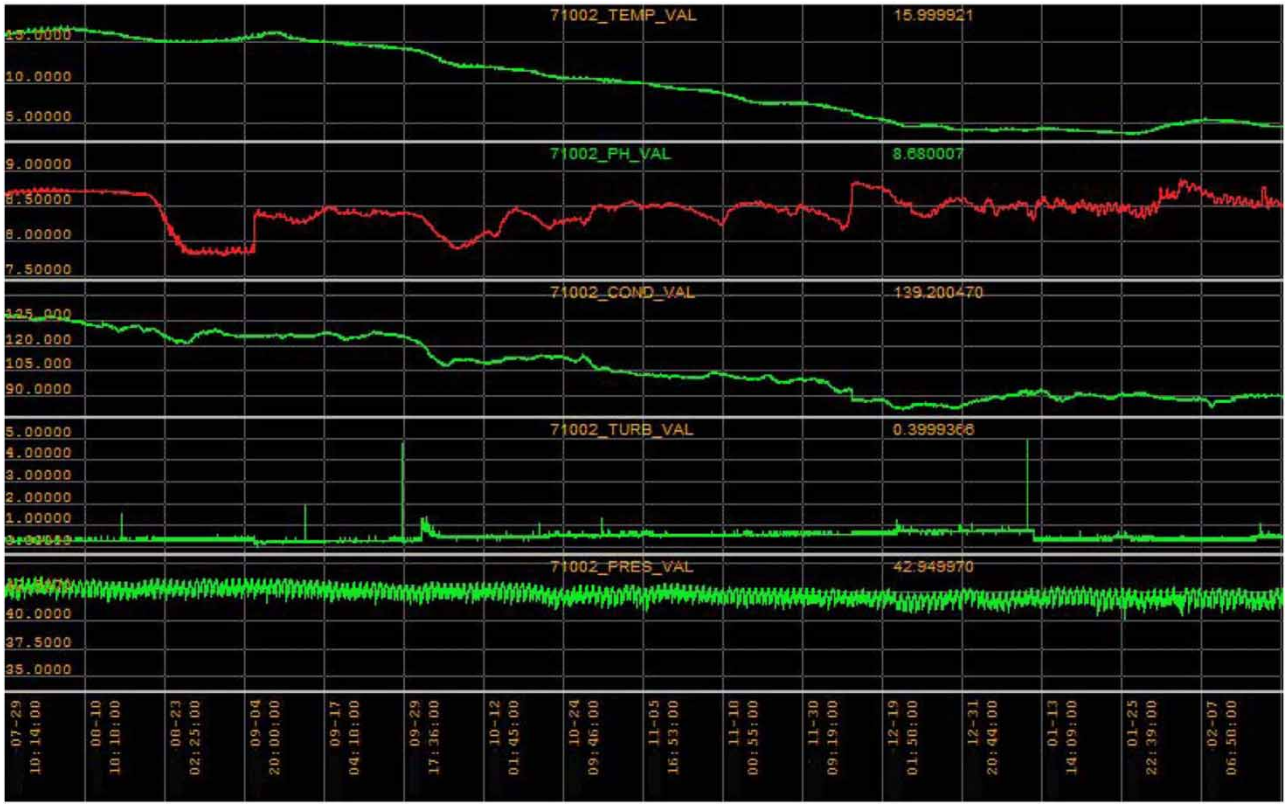

Figure 2 | Signal Data Explorer opening a water quality sensor data file.

variable assigned to a bin and each bin corresponding to a different bit that will be set in the binary pattern – with only one bit set (as illustrated as part of Figure 3). The simplest binning method is to use bins of equal width, although AURA allows the option to set a threshold for the number of values that can be placed in the extreme bins and, once exceeded, the bin values are reset and the memory retrained. The codes for each variable are then concatenated to create a binary representation of the state, which is stored in a binary CMM (Willshaw *et al.* 1969) of an AURA associative memory. A binary CMM is a single layer, fully connected network that is capable of very fast storage and retrieval of data. A CMM with input width n and output width m can be represented as a $n \times m$ binary matrix $\mathbf{M}$. For a given input binary vector $\mathbf{I}_k$ and associated binary output vector $\mathbf{O}_k$, the kth training update of a CMM is defined as:

$$\mathbf{M}_k = \mathbf{M}_{k-1} \cup \mathbf{I}_k^T \mathbf{O}_k \tag{9}$$

where $\mathbf{M}_k$ and $\mathbf{M}_{k\text{-}1}$ are the CMM after and before the training (with $\mathbf{M}_0 = 0$ and $\cup$ denoting a logical OR operation between the vectors). The recall vector $\mathbf{S}_i$ associated to the input $\mathbf{I}_i$ is defined as:

$$\mathbf{S}_i = \mathbf{I}_i \mathbf{M} \tag{10}$$

This recall vector is generally an integer vector and the value of each element of the recall vector is called the 'score' of the CMM matching on the relevant column vector. The recall vector can then be thresholded to a binary output vector by either using a fixed threshold or selecting the L closest matches. This process is shown in Figure 3 using small vectors for illustrative purposes (fixed threshold of 2).

In practice, the recall system needs to factor in not only the number of bins that match exactly, but also the distance between the assigned bins when they differ since this will provide important information on the closeness of match.

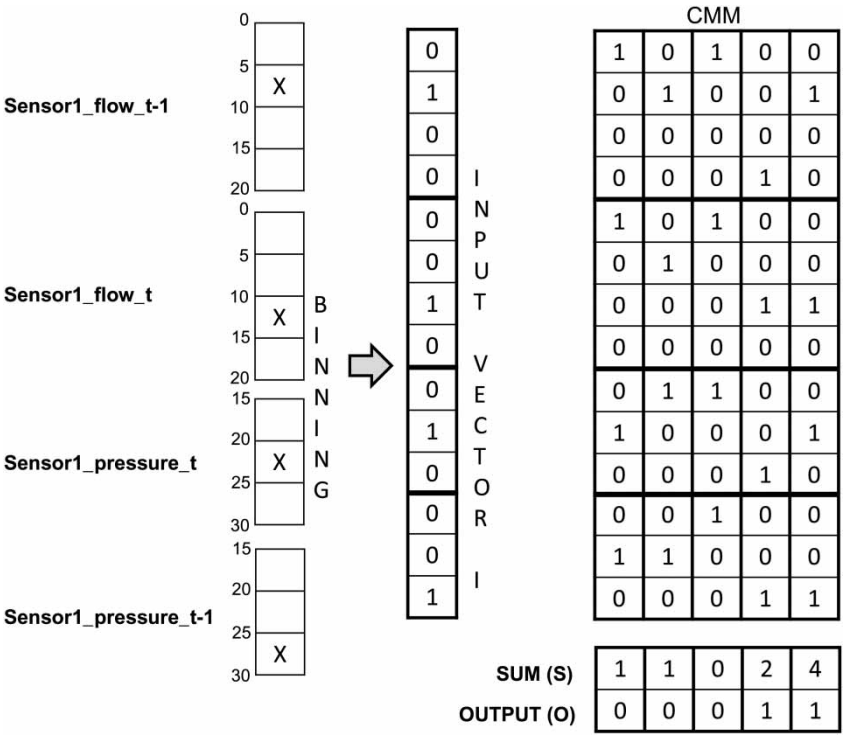

Figure 3 | Input variable binning and the CMM recall process.

This can be achieved by weighting the inputs. Hence AURA scores can be used to apply weights to the bins, according to their distance from the current values. A vector of scores (rather than a binary pattern) is created, which defines a set of kernels that quantify the distance of each bin from the value. For example this could be a triangular shaped kernel centred around the recall value. Weeks *et al.* (2003) discuss various kernels that can be used to provide different approximations of distance. However, a kernel which approximates to the Euclidean distance between two points is described in Equation (11) and has been found to provide good results for fixed binning (Hodge *et al.* 2004).

$$\text{Parabolic}_{\text{bins}_k} = \left[\left(\frac{\max(n)}{2}\right)^2 - (\text{bins}_t - \text{bins}_k)^2\left(\frac{\max(n)^2}{n_f^2}\right)\right] \tag{11}$$

where the output is the value of bin number k (bins_k) and the value of the variable has been assigned to bin t (bins_t), $\max(n)$ is the number of bins for any variable and n_f is the number of bins for this variable.

This scoring system now more accurately reflects the actual distance between each stored point and the recalled vector and hence the current and recorded asset states. Some modifications are then necessary to the threshold technique (Austin *et al.* 2010). AURA k-Nearest Neighbour (k-NN) can then be used as a filter to reduce a large number of stored states to a more manageable quantity of closer matches (Hodge & Austin 2005).

The k-NN pattern matching method is widely used in data clustering, classification and prediction. Based on a specific distance metric or similarity measure, k-NN examines vector distances to determine the nearest neighbours (Cover & Hart 1967). One approach could be to use k-NN on the raw values of each variable at each time interval and compare the recalled points to each of these. However, the resultant time complexity of performing these separate comparisons would severely limit the number of states that can be stored in the system. The standard algorithm is computationally slow for large datasets. A binary neural network-based k-NN has been developed (Zhou *et al.* 1999; Hodge & Austin 2005) which can search millions of

states very quickly. AURA k-NN is efficient and scalable, and has shown to be up to four times faster than the traditional k-NN (Hodge *et al.* 2004). AURA Alert is the implementation of AURA within the SDE.

Using this approach, the system will locate the nearest k matching patterns. The score associated with the most closely matching column(s) can then be used to determine how different the current state is to any that have been seen before, to provide a measure of the novelty of the event. The AURA Alert software can thus continuously provide a measure of novelty across a time series. Note that only an outline of CMMs and AURA Alert has been provided here. An in-depth description of CMMs, AURA and AURA k-NN can be found in: Austin (1995), Zhou *et al.* (1999), Liang & Austin (2003), Hodge *et al.* (2004), Hodge & Austin (2005), Furber *et al.* (2007) and Fletcher *et al.* (2008).

Water distribution system time series data

Data streams from WDSs can be somewhat different to other domains such as found in engine or power plant monitoring. Some variables, particularly hydraulic parameters, such as flow and pressure, possess a diurnal pattern which reflects the daily demand profile dominated by residential use, pressure in Figure 2 illustrates this. Some water quality measurements also reflect this, so that chlorine concentration for example will (generally) have a periodic sinusoidal like profile. However, other water quality parameters such as conductivity are more similar to those measurements usually encountered in CM. Finally, some can have both characteristics, such as turbidity (as seen in Figure 2).

In order to use AURA alerts on data with periodic (e.g. daily) cycles, it is necessary to introduce an extra 'time of day' variable (e.g. the number of elapsed hours of the day). This enables AURA alerts to detect patterns in the data that are unusual at that time of day. The data collected from sensors are first formatted into input files for a MATLAB pre-processing program which identifies and fills in any missing timestamps or values so as to provide a continuous stream of data. The data are finally reformatted into an appropriate comma delimited format required by the SDE. Note though that for non-periodic data streams, the AURA system is able to deal with completely missing data, with a zero code indicating the absence of data – particularly useful for dealing with instrumentation or telemetry problems in online systems.

AURA Alert is then provided with data from an extended period of time (at least 2 weeks) during which the WDS sensor has been known to perform correctly with 'normal' conditions in the distribution system. At regular time intervals during this period, the values of a representative set of variables from the data are converted into a pattern, which represents the state of the WDS zone at that time instance. This pattern is then stored in an AURA associative memory.

RESULTS AND DISCUSSION

Pattern matching

Data analysis was conducted using the SDE and a query by content approach for pattern matching. In addition, pattern matching software was developed in C# using Microsoft Visual Studio. Libraries of event profiles were created from .csv data to allow batch processing. Ten DMA inlet flows (A to J) were obtained for a large water supply system, with a mixture of urban and rural areas, for an approximate 8 month period for use in selecting burst profiles (industry standard 15 minute sampled data) along with the Work Management System (WMS) mains repairs record. A pattern library of known bursts for these DMA flow inlets using the SDE was assembled from this dataset. These were identified from within the 10 DMA flow inlet datasets (normalised as described previously in order to allow generalisation from the DMA flow values) by using WMS information to confirm large burst events and hence creating a set of profiles consisting of a number of consecutive measurements (described in the pattern matching section). These were chosen to capture the significant first features of change in parameter due to an event – using between one and two hours of data. The SDE allows searching for similar patterns in this library. One DMA (G) was held back for testing using the pattern library. An example is provided in Figure 4 of a detected burst in this DMA, which was matched with a very high probability to a burst from another DMA.

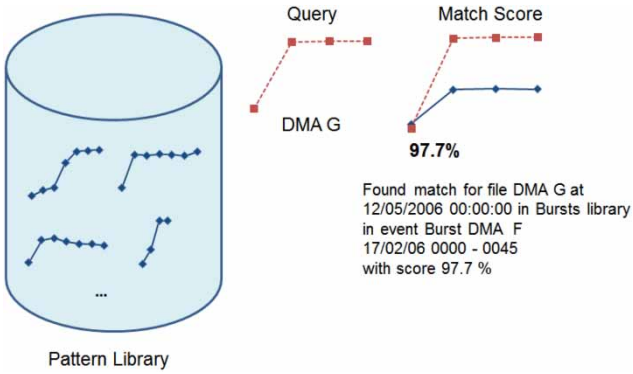

Figure 4 | Pattern matching bursts in DMA flow data.

Three other matches of above 90% match score were obtained for DMA G for the whole period of analysis – summarised in Table 2. In the case of each of the detections, visual analysis revealed that the profile was briefly unusual, although there was only one correlation found with WMS history. The results for an Artificial Intelligence (AI) analysis system and all mains burst repairs (MR) for the same DMA are also reported in Table 2 (after Mounce *et al.* 2007). Table 2 reports three MR in the period of which two were detected by the AI system (the other having no significant impact on the nightline). In particular, a burst was repaired on 24/12/05 of significant duration detected by the AI system (a total of three AI detections > 85% confidence) but not resulting in a hit using pattern matching.

Although this example illustrates the transferability of the concept of a 'burst' pattern, a limitation in the approach is in the manual assembly of the pattern library and the uncertainty prevalent in defining event classes for WDS. Even when limited to burst only patterns, performance on the test DMA was found not as accurate as an AI system utilising outlier detection. Accurate selection of precursor patterns is also far from obvious. Using AURA Alert to automatically calculate a novelty score for any type of event, possibly never encountered before, was thus identified as a more attractive technique with the possibility of detecting precursor features before major failure.

AURA Alert

The AURA Alert system utilising CMMs has been used for the detection of irregularities in highly complex assets in a variety of different industries. Applying AURA Alert on real data from two different WDSs to explore the capabilities of the method and results obtained are now described.

Flow data analysis

The DMA inlet flows A – J, used in the pattern matching test, were each analysed by the AURA system and performance compared to WMS and the aforementioned AI system (a 4 week period was used to create the CMM model).

Table 2 | Pattern matching results for DMA G compared to AI system

Pattern matching (PM) positive classification period	WMS record repair date	AI positive classification period
None	MR 24/12/05	23/12/05 14:00–24/12/05 06:15 (96% confidence, burst est. 0.74)
Match at 15/04/2006 07:30 in PM library with Burst DMA E flow 190406 0000-0130 (score 91%)	None	None > 85% confidence
Match at 16/04/2006 06:00 in PM library with Burst DMA E flow 190406 0000-0130 (score 93%)	None	None > 85% confidence
None	MR 09/02/06	None >85% confidence
Match at 12/05/2006 00:00 in PM library with Burst DMA F flow 170206 0000-0045 (score 98%)	MR 12/05/06	11/05/06 21:15–12/05/06 11:45 (99% confidence, burst est. 3.13)
Match at 29/05/2006 08:45 in PM library in with Burst DMA E flow 190406 0000-0130 (score 91%)	None	None >85% confidence
None	None	03/06/06 14:30–04/06/06 05:30 (99% confidence, burst est. 0.60)

A match score threshold of 85% was used to identify reasonably large deviations from normality resulting in 20 overall detections (in comparison to 16 for the AI system). Of these, four detections corresponded well to WMS burst repairs (for the AI based system this number was five – with three of these detected by both systems). Of the remainder, 13 were correlated visually with abnormal temporary increases in flow and three with likely short sensor drop outs to zero.

Overall the performance was thus comparable to the AI detection as reported in Mounce *et al.* (2007). AURA offers other possible advantages such as across multi-parameter analysis or potential short precursor event detection as further explored in the next two examples.

Water quality example

A multi-parameter water quality dataset was obtained for a measuring instrument based at a DMA inlet in an urban WDS deployed as part of a pilot study. Parameters measured were water temperature, pH, conductivity, turbidity and pressure at a 5 minute resolution. Data from a period of several weeks when the DMA was considered to be operating normally were presented to the AURA Alert system and the learned configurations encountered were stored in the AURA memory. Figure 5 shows the five channels corresponding to the raw data. The AURA Alert output can be seen in the 'Match Strength' channel (bottom axes), which has a value of 100 when in a previously seen state and drops down when a novelty is detected. In Figure 5 note how the matching strength remains high during the period of normal activity earlier in the period (the greyed out section indicating the end of the training data) but later reports the presence of novelty, indicating that the asset state has departed from its usual operating behaviour (in fact this was a known burst affecting the DMA being monitored).

In addition to reporting the matching strength of the state of the system at each time instance, AURA Alert is able to indicate which channels are the likely causes of the irregularities. By using an L-Max threshold (the metric where L highest sums are set to 1 and all others to 0) on the AURA output, the most similar stored pattern to the current asset state can be obtained. By comparing the current state to the most similar state, the causes of the differences can be calculated and reported along with the matching strength. The reported novelties (not shown here) indicate that the turbidity, pressure and conductivity are deviating from their expected values, suggesting a burst event. Although a burst has been used for illustration, the proliferation in measured parameter options in new WDS water quality instrumentation paves the way for detection and classification (based on which parameters are novel) for other types of abnormality such as contamination events (intentional or accidental) (e.g. Leeder *et al.* 2012).

Pre-cursor example

The final example presented is in the use of AURA Alert to identify novelties in multiparameter data several days before a catastrophic failure in a complex asset, without any prior knowledge of similar failures. A flow and pressure dataset was assembled for a DMA. The data consisted of 15 minute readings, the WMS record and any associated customer contacts (CC) (complaints to call centres). These data include pressure data from the DMA inlet in addition to two specific point pressure loggers located at critical (determined by expert judgement) locations in the DMA. Hydraulic data were utilised, with AURA trained using several weeks of normal data, and a test period with known multiple events and supporting information has then been used to illustrate the possibility of precursor detection. Figure 6 provides the Match Strength output with WMS and CC information overlaid and in addition the online detection from an online AI system (Mounce & Boxall 2010).

Figure 6 plots a period of 9 days during which two water main burst repairs were flagged and marked as repaired in the DMA. Information in the WMS reveals that one repair had a start date of 20th September and a completion on 24th September. The second had a start date of 24th September and a completion date the day after. We see from the flow plot, that a burst main repair resulted in a drop of the nightline on 23rd September once completed. Before this, a new burst was first detected by an online AI system 06:00 23/9, which preceded a number of CCs (11 customers complaining of no water, and two of discolouration). The Match Strength drops below 90 several hours before this. However, of more interest is the large drop in Match Strength on 18th September

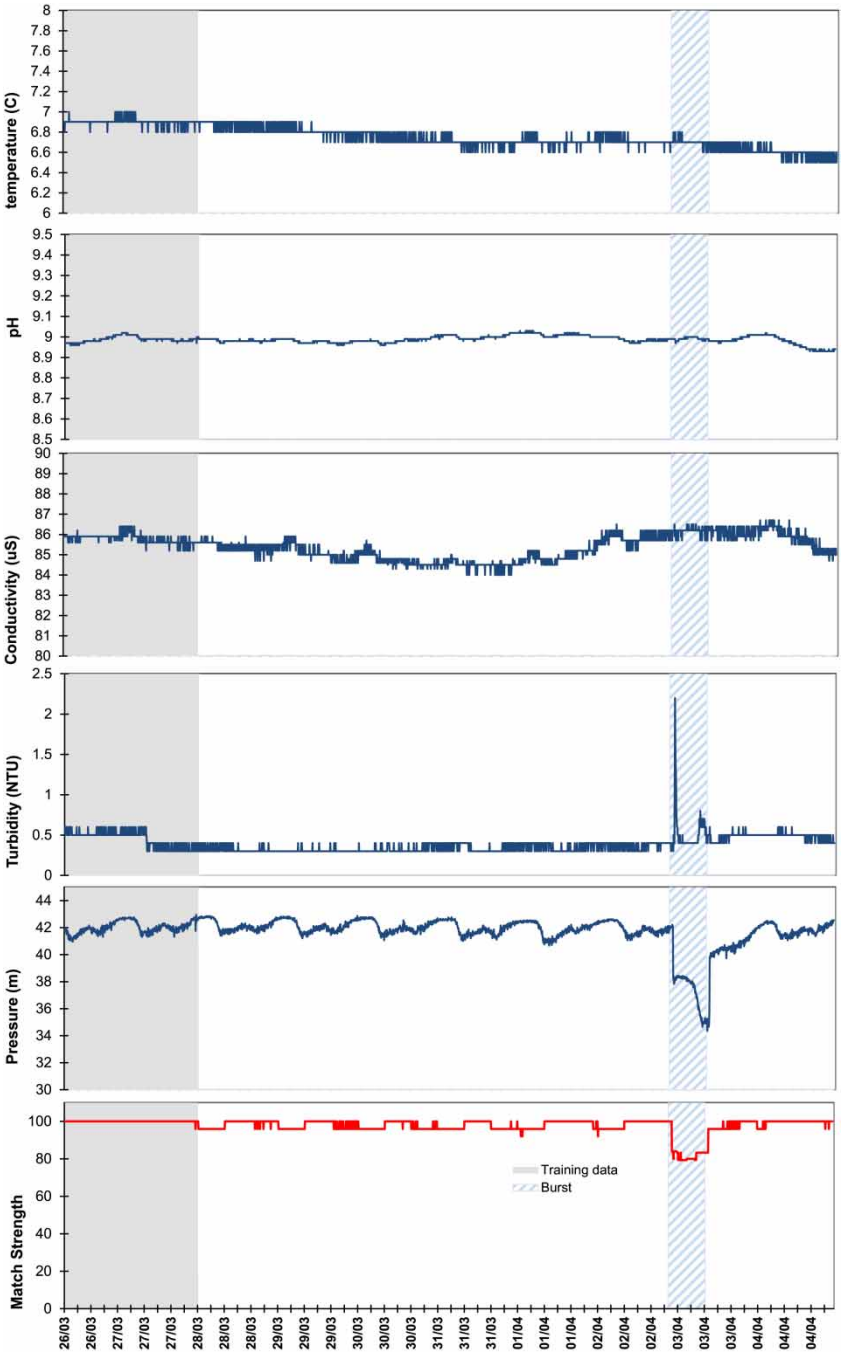

Figure 5 | AURA Alert for abnormal event in water quality data.

around midnight prior to any other warning of a problem and corresponding to short duration drops in both pressure and flow; this may be a burst precursor indicative of a developing problem or some activity on the network which subsequently caused the major burst several days later. The fact that the water company noted a repair start date to the WMS database on 20th September supports this. Whilst Figure 6 shows the potential for precursor detection, confirmation can be rather subjective due to the resolution of data and in particular the

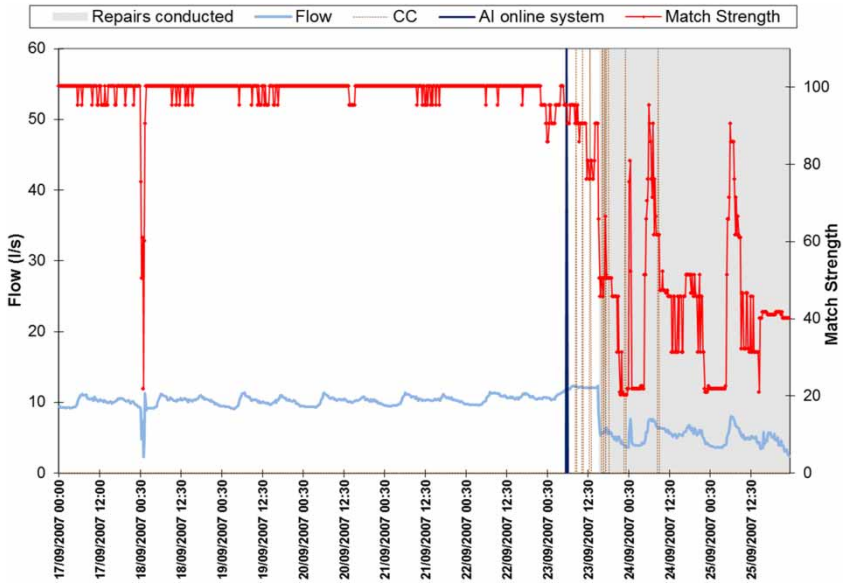

Figure 6 | AURA Alert Match Strength and supporting information for potential precursor.

supporting information. Validation of this capability would require a more extensive bespoke study.

It has been demonstrated that the AURA Alert system has the potential for detecting changes across multiple parameters, allowing robust detection and information for interpretation, and offering scope for detection of event precursors. Timely event detection and diagnosis offer significant improvements in service delivery with a move towards proactive maintenance, while the implication of precursor information is to provide network engineers additional time to investigate the cause of abnormal conditions and perhaps prevent major asset failure before customers are impacted. Of course, datasets with more exhaustive information (such as known artificial hydrant flushing) could be used to evaluate more rigorous quantifiable error metrics such as the level of false-positives.

General discussion

WDS sensors monitor assets (reservoirs, pipes, valves, etc.) with the performance of these assets being indicated by the collected measurements. At the present time, the granularity, i.e. number of devices and sampling interval, is quite limited compared to other industries. However, the quantity and complexity of sensor and environmental data are growing at an increasing rate and it seems clear that in the future the water sector can and should be penetrated by Information and Communications Technologies and Internet-like technologies. It is easy to anticipate that the environment may before very long be teeming with tens of thousands of small, low-power, wireless sensors. Each of these devices will produce a stream of data, and those streams will need to be monitored and combined to detect changes of interest in the environment. The easier it is to collect and analyse large datasets the more water utilities will collect and, in a decade, tens or even hundreds of petabytes of data may be routinely available. Demands for solutions and tools will become more urgent to meet the aspiration for intelligent water networks, proactively managed through access to timely information. Permanent installation of high frequency (several hundred or even thousand Hz) pressure monitoring devices may also become routine and pilot studies using these have demonstrated how the arrival times of the burst induced wave at the measurement points can be used to derive the location of the burst using transients (Misiunas *et al.* 2005). The data compression facilities of systems such as AURA could prove very useful for these future data quantities.

This proliferation of monitoring will facilitate the continuous and simultaneous monitoring of the complete WDS (or at least significant sub-areas). By evaluating deviation from normality from a set of distributed sensors, both detection and location of abnormal events will be

possible – such as using multiple pressure loggers to locate a burst (e.g. Farley *et al.* 2013; Romano *et al.* 2013). The monitoring of sewerage systems has not progressed as far as for WDSs, however there is increasing interest and deployment of instrumentation for example for Combined Sewer Overflow level measurement and pump station flows. There are many other potential applications in the water resources domain.

AURA Alert is being developed as an online system which automates the training data selection (by selecting data with high Match Strength) and use of validation data for selecting Match Strength thresholds for alert generation. This can be deployed by the cloud as a Software as a Service, leveraging Grid technology and permitting secure, rapid delivery of information to the viewer.

CONCLUSIONS

The effective and efficient operation of WDSs is essential for three important reasons: maintaining safe and continuous supply to consumers, avoiding loss of water resources through leaks and bursts in the pipe network, and reducing the energy and other resources input to the system and so minimising the carbon footprint of water system operations. To achieve this efficiency, information is continually required about current system performance, so adjustments can be made where necessary and interventions can occur before any fault or failure impacts on the customer. This paper has presented the use of pattern matching and binary associative neural networks using time series from WDS. Using AURA Alert, time series data from sensors (variables) are converted into vectors using a quantisation process. Vectors are then stored in a historical database in the correlation matrix memory. New data presented as vectors can either be used to generate the k best matching historical patterns or alternatively a measure of novelty (termed Match Strength) can be generated. One of the major features of the system is its ability to search small and very large datasets very quickly. The key conclusions of this research are as follows:

- A pattern matching approach can be proficient at finding known patterns in data and has been applied successfully for many applications. The transferability (i.e. not tuned per DMA) of burst patterns was demonstrated here to some extent. However, overall the performance was found to be not as high as when using outlier detection based methods for this type of WDS time series data. A limitation of the approach is in the manual assembly of the pattern library and the uncertainty prevalent in defining event classes for WDS.
- AURA Alert (Advanced Uncertain Reasoning Architecture, utilising a class of binary neural network built on CMMs) can rapidly learn and model the normal operating envelope for a system, with the ability to search through complex high-dimensional multivariate spaces to detect deviations from normal conditions. The novel use of AURA Alert in WDS so as to automatically calculate a continuous novelty score for every time step and hence enable the detection of any type of event, possibly never encountered before, was proposed, explored and demonstrated. Examples have demonstrated successful early detection of abnormality in systems using multi-parameter data as well as significant potential for precursor event detection beyond typical outlier detection approaches. These precursors could be linked to appropriate maintenance requirements for water infrastructure.

ACKNOWLEDGEMENTS

This work was supported by the Pennine Water Group – Urban Water Systems for a Changing World Platform Grant (EP/I029346/1) and by the Pipe Dreams project (EP/G029946/1), both funded by the UK Science and Engineering Research Council. The authors would like to thank Yorkshire Water Services for data provision.

REFERENCES

Aisopou, A., Stoianov, I. & Graham, N. 2012 In-pipe water quality monitoring in water supply systems under steady and unsteady state flow conditions: a quantitative assessment. *Water Research* **46**, 235–246.

Akselaa, K., Akselab, M. & Vahalaa, R. 2009 Leakage detection in a real distribution network using a SOM. *Urban Water* **6** (4), 279–289.

Austin, J. 1995 Distributed associative memories for high speed symbolic reasoning. *International Journal on Fuzzy Sets and Systems* **82**, 223–233.

Austin, J., Brewer, G., Jackson, T. & Hodge, V. J. 2010 AURA-Alert: The use of binary associative memories for condition monitoring applications. In: *Proceedings of 7th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies: (CM 2010 and MFPT 2010)*, vol. 1, pp. 699–711.

Bishop, C. M. 2007 *Pattern Recognition and Machine Learning*. Springer, New York.

Branisavljević, N., Prodanović, D. & Pavlović, D. 2010 Automatic, semi-automatic and manual validation of urban drainage data. *Water Science and Technology* **62** (5), 1013–1021.

Cover, T. & Hart, P. 1967 Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13** (1), 21–27.

Farley, B., Mounce, S. R. & Boxall, J. B. 2013 Development and field validation of a burst localization methodology. *ASCE Journal of Water Resources Planning and Management* **139** (6), 604–613.

Fletcher, M., Liang, B., Smith, L., Knowles, A., Jackson, T., Jessop, M. & Austin, J. 2008 Neural network based pattern matching and spike detection tools and services in the CARMEN neuroinformatics project. *Neural Networks* **21**, 1076–1084.

Furber, S. B., Brown, G., Bose, J., Cumpstey, J. M., Marshall, P. & Shapiro, J. L. 2007 Sparse distributed memory using rank-order neural codes. *IEEE Transactions on Neural Networks* **18** (3), 648–659.

Hodge, V. & Austin, J. 2004 A survey of outlier detection methodologies. *Artificial Intelligence Review* **22**, 85–126.

Hodge, V. & Austin, J. 2005 A binary neural k-Nearest Neighbour technique. *Knowledge and Information Systems* **8** (3), 276–292.

Hodge, V., Lees, K. J. & Austin, J. 2004 A high performance *k*-NN approach using binary neural networks. *Neural Networks* **17** (3), 441–458.

Hopfield, J. J. 1982 Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **81** (10), 3088–3092.

Jarrett, R., Robinson, G. & O'Halloran, R. 2006 On-line monitoring of water distribution systems: data processing and anomaly detection. In: *Proceedings of the 8th Water Distribution System Analysis Symposium*, Cincinnati, USA, August 27–30.

Keogh, E., Lonardi, S. & Chiu, W. 2002 Finding surprising patterns in a time series database in linear time and space. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 23–26, 2002, Edmonton, Alberta, Canada, pp. 550–556.

Keogh, E. & Smyth, P. 1997 A probabilistic approach to fast pattern matching in time series databases. In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining – KDD 97*, pp. 24–30.

Kosko, B. 1988 Bidirectional associative memories. *IEEE Transactions On Systems, Man, And Cybernetics* **18** (1), 49–60.

Leeder, A., Mounce, S. R. & Boxall, J. B. 2012 Analysis of multi-parameter water quality data using event detection software on laboratory simulated events. In: *Proceedings of WDSA 2012 Conference*, Australia.

Liang, B. & Austin, J. 2003 Improved high performance k-NN classifier using a binary neural network. In: *Eighth International Conference on Engineering Applications of Neural Networks (EANN-03), Spain*, pp. 148–153.

Masters, T. 1993 *Practical Neural Network Recipes in C++*. Academic Press, San Diego, CA.

McKenna, S. A., Hart, D., Klise, K., Cruz, V. & Wilson, M. 2007 Event detection from water quality time series. In: *Proceedings of World Environmental and Water Resources Congress*, ASCE, Reston, VA.

Misiunas, D., Vítkovský, J., Olsson, G., Simpson, A. & Lambert, M. 2005 Pipeline burst detection and location using a continuous monitoring of transients. *Journal of Water Resources Planning and Management, ASCE* **131** (4), 316–325.

Mounce, R. B., Hollier, G., Smith, M., Hodge, V. J., Jackson, T. & Austin, J. 2013 A metric for pattern-matching applications to traffic management. *Transportation Research Part C: Emerging Technologies* **29**, 148–155.

Mounce, S. R. & Boxall, J. B. 2010 Implementation of an on-line artificial intelligence district meter area flow meter data analysis system for abnormality detection: a case study. *Water Science and Technology: Water Supply* **10** (3), 437–444.

Mounce, S. R., Boxall, J. B. & Machell, J. 2007 An artificial neural network/fuzzy logic system for DMA flow meter data analysis providing burst identification and size estimation. In: *Water Management Challenges in Global Change* (B. Ulanicki, K. Vairavamoorthy, D. Butler, R. Bounds & F. Memon, eds). Taylor and Francis, London, pp. 313–320.

Mounce, S. R., Boxall, J. B. & Machell, J. 2010 Development and verification of an online artificial intelligence system for burst detection in water distribution systems. *Journal of Water Resources Planning and Management* **136** (3), 309–318.

Mounce, S. R. & Machell, J. 2006 Burst detection using hydraulic data from water distribution systems with artificial neural networks. *Urban Water Journal* **3** (1), 21–31.

Mounce, S. R., Mounce, R. B. & Boxall, J. B. 2011 Novelty detection for time series data analysis in water distribution systems using Support Vector Machines. *Journal of Hydroinformatics* **13** (4), 672–686.

Olsson, G. 2012 *Water and Energy: Threats and Opportunities*. IWA Publishing, London.

Palau, C. V., Arregui, F. J. & Carlos, M. 2012 Burst detection in water networks using principal component analysis. *Journal of Water Resources Planning and Management* **138** (1), 47–54.

Parliamentary Office of Science and Technology 2011 Water in Production and Products. POSTnote 345.

Romano, M., Kapelan, Z. & Savić, D. A. 2013 Geostatistical techniques for approximate location of pipe burst events in

water distribution systems. *Journal of Hydroinformatics* **15** (3), 634–651.

Romano, M., Kapelan, Z. & Savić, D. A. 2014 Automated detection of pipe bursts and other events in water distribution systems. *Journal of Water Resources Planning and Management* **140** (4), 457–467. Available from: http://ascelibrary.org/doi/abs/10.1061/(ASCE)WR.1943-5452.0000339.

Solomatine, D. 2002 Data-driven modelling: paradigm, methods, experiences. In: *Proc. 5th International Conference on Hydroinformatics*. pp. 1–5. http://www.unesco-ihe.org/hi/hi/staff/sol/HI2002-DDM-76transp.pdf.

Weeks, M., Hodge, V., O'Keefe, S., Austin, J. & Lees, K. 2003 Improved AURA k-Nearest Neighbor approach. In: *Proceedings of IWANN–2003, International Work-conference on Artificial and Natural Neural Networks, Mahon, Menorca, Balearic Islands, Spain. Lecture Notes in Computer Science (LNCS) 2687*, Springer Verlag, Berlin.

Willshaw, D. J., Buneman, O. P. & Longuet-Higgins, H. C. 1969 Non-holographic associative memory. *Nature* **222**, 960–962.

Wu, Z. Y., Farley, M., Turtle, D., Kapelan, Z., Boxall, J. B., Mounce, S. R., Dahasahasra, S., Mulay, M. & Kleiner, Y. 2011 *Water Loss Reduction* (Z. Wu, ed.). Bentley Systems, Exton, PA.

Xu, D. L., Liu, J., Yang, J. B., Liu, G. P., Wang, J., Jenkinson, I. & Ren, J. 2007 Inference and learning methodology of belief-rule-based expert system for pipeline leak detection. *Expert Systems with Applications* **32**, 103–113.

Ye, G. & Fenner, R. 2010 Kalman filtering of hydraulic measurements for burst detection in water distribution systems. *ASCE Journal of Pipeline Systems Engineering and Practice* **2** (1), 14–22.

Zhang, Z., Jiang, J., Liu, X., Lau, R., Wang, H. & Zhang, R. 2010 Real time hybrid pattern matching scheme for stock time series. *ADC '10 Proceedings of the Twenty-First Australasian Conference on Database Technologies* **104**, 161–170.

Zhou, P., Austin, J. & Kennedy, J. 1999 Chapter 9, Online monitoring and detection. A high performance k-NN classifier using a binary correlation matrix memory (M. J. Kearns, S. A. Solla & D. A. Cohn eds). In: *Advances in Neural Information Processing Systems, Vol*. 11. MIT Press, CA.

First received 13 May 2013; accepted in revised form 13 August 2013. Available online 8 October 2013