sciendo

Int. J. Appl. Math. Comput. Sci., 2018, Vol. 28, No. 2, 283–295
DOI: 10.2478/amcs-2018-0020

amcs

AN UNSUPERVISED APPROACH TO LEAK DETECTION AND LOCATION IN WATER DISTRIBUTION NETWORKS

MARCOS QUIÑONES-GRUEIRO [a,*], CRISTINA VERDE [b], ALBERTO PRIETO-MORENO [a], ORESTES LLANES-SANTIAGO [a]

[a]Department of Automation and Computing
Havana University of Technologies José Antonio Echeverría (CUJAE)
114, e/ Ciclovía y Rotonda, Marianao, 19390, La Habana, Cuba
e-mail: marcosqg@automatica.cujae.edu.cu

[b]Institute of Engineering
National Autonomous University of Mexico (UNAM)
Coyoacán, 04510 México DF, Mexico

The water loss detection and location problem has received great attention in recent years. In particular, data-driven methods have shown very promising results mainly because they can deal with uncertain data and the variability of models better than model-based methods. The main contribution of this work is an unsupervised approach to leak detection and location in water distribution networks. This approach is based on a zone division of the network, and it only requires data from a normal operation scenario of the pipe network. The proposition combines a periodic transformation and a data vector extension together with principal component analysis of leak detection. A reconstruction-based contribution index is used for determining the leak zone location. The Hanoi distribution network is employed as the case study for illustrating the feasibility of the proposal. Single leaks are emulated with varying outflow magnitudes at all nodes that represent less than 2.5% of the total demand of the network and between 3% and 25% of the node's demand. All leaks can be detected within the time interval of a day, and the average classification rate obtained is 85.28% by using only data from three pressure sensors.

Keywords: water distribution networks, leak location, unsupervised methods, principal component analysis, demand model.

1. Introduction

Water resources are one of the most important assets in modern society. Water distribution networks (WDNs) deliver drinking water to different types of consumers all around the world (Łangowski and Brdys, 2017). WDNs are non-linear dynamic systems formed by supply systems, pipe networks and water control elements and are governed by main physical laws, a system layout and consumer demand.

As pipe networks degrade because of system ageing and other phenomena, distribution systems experience different kind of faults. Among them, damage to the network infrastructure provokes pipe bursts and leaks that cause water losses with severe economic and environmental consequences to the water companies (Colombo and Kamey, 2002). Service disruptions represent a risk to consumer health because of the possible water scarcity. Moreover, water losses have a negative impact on the water companies' operational performance, customer service and reputation (Romano *et al*., 2013).

Water loss detection and location strategies are usually based on inference methods, which build models to represent the behavior of district metered areas (DMAs) by using measurement data from permanently installed sensors. These tools are cost-effective, noninvasive and do not require a survey of the whole DMA by trained personnel (Romano *et al*., 2010). The basic idea behind inference-based fault detection and isolation is to make a decision about the DMA behavior by comparing the measurements with the model output signals. Inference

*Corresponding author

methods can be divided into two types: analytic and data-driven.

Analytic approaches (often referred to as model-based) use fault detection and isolation (FDI) methods by developing an analytic model of the DMA based on the main physical laws describing its operation. The identification of an analytical model is not an easy task for real DMAs because of the amount of parameters, constraints and uncertainties involved.

Data-driven approaches adopt a pattern recognition philosophy by assuming that the faults affect some observable system variables and that historical data are available. These methods have found great applicability in real-life DMAs mainly because they do not need detailed knowledge about the pipe network parameters or system layout (Laucelli *et al*., 2016). Furthermore, in recent years an increasing amount of historical data from WDNs is available given the widespread use of modern instrumentation together with supervisory control and data acquisition (SCADA) systems. Hence, in the last 10 years various data-driven FDI strategies have been proposed to cope with the water loss detection and location.

Neural networks were proposed for water loss detection in DMAs. Self-organizing maps, feedforward multilayer perceptrons and binary associative neural networks were used for leak detection (Aksela *et al*., 2009; Mounce *et al*., 2014). Statistical and signal analysis tools were developed for the monitoring of DMAs. Principal component analysis, evolutionary polynomial regression and clustering based outlier analysis were used for this purpose (Nowicki *et al*., 2012; Palau *et al*., 2012; Laucelli *et al*., 2016; Wu *et al*., 2016). These unsupervised approaches do not allow the leak location.

Recently, a mixed model-based/data-based approach has been proposed for water loss detection and location (Wachla *et al.*, 2015; Soldevila *et al.*, 2016; 2017; Zhang *et al.*, 2016; Moczulski *et al.* , 2016). These works suggest the use of a hydraulic model of the DMA for generating data sets which are then used for the calibration of a pattern recognition method. The task of this method is on-line fault detection and location. These supervised approaches require historical sets of normal and fault data for performing the leak detection and location tasks.

Zhang *et al*. (2016) and Moczulski *et al*. (2016) consider the fault location issue as a fault zone location problem. The zone can be defined as a network subarea formed by a set of physically interconnected nodes such that a DMA is divided into different nonoverlapping zones. Thus, once a fault has been detected, the fault zone location problem is formulated as the problem of identifying the zone where a fault is occurring, i.e., a fault is geographically isolated from the rest of network zones. In a data-driven framework, the data sets obtained by simulating the network with specific operation conditions and a single fault (leakage) with a specific magnitude, which can occur at different locations (nodes) within a zone, form a class. Therefore, each class is uniquely associated with a zone, and multiple classes form the training and validation data from a pattern recognition perspective. Supervised tools have then been applied for fault zone location, e.g., neuro-fuzzy classifiers and multiclass support vector machines. The main advantage of this approach over the model-based methods is that the network uncertainty can be handled while obtaining a satisfactory FDI performance.

A data set of each fault scenario (leakage location) is characterized by the network operation conditions, such as demand pattern variability, measurement noise, and the fault features (pipe break size, time of occurrence). It is then assumed that the data sets of each class represent the most probable conditions for the fault scenario. Therefore, a vast amount of data must be collected for each class given the range of possible scenarios, and if the real leakage features and network conditions differ from the simulated ones, then the fault location performance may deteriorate. Moreover, in these works the demand uncertainty is not explicitly considered, and some of them assume that multiple flow sensors are installed on the DMA (Wachla *et al*., 2015; Moczulski *et al*., 2016). Ultimately, obtaining a data set of all the possible leak scenarios is not a feasible task even for networks with a model that is available. The drawbacks mentioned above motivate this paper, and the main contribution is a fault detection and location approach that considers an unsupervised method. Moreover, only pressure sensors are used, because, as indicated by Jung and Lansey (2015), they have lower costs than flow sensors, and they can be easily installed and maintained.

The zone division can be performed according to many different criteria such that each zone is representative of a demand pattern. This idea has been adopted by Sanz *et al*. (2015) such that a model of the network and a demand calibration process are used for establishing zones and node membership to each zone. When a leak occurs, the recalibration of the demands allows determining a group of candidate nodes. This approach relies, however, on the availability of an explicit model of the network with parameters that are difficult to estimate for real networks. Therefore, in this work it is considered that the pressure sensors have already been installed by following the experts' knowledge, and the zone division is performed geographically according to the areas where the sensors are placed such that there is a one-to-one correspondence.

The unsupervised method combines a preprocessing procedure with a historical data-based tool such as principal component analysis (PCA). The main problem of using traditional data-based tools such as PCA for fault detection in WDNs is that the assumption of weakly

stationary signals is not generally satisfied. A previous work deals with this issue by using the information of a segmented demand pattern for building multiple PCA models (Quiñones-Grueiro *et al*., 2016). Thus, each model is employed to perform the fault detection task within a specific time interval where variables are considered to be stationary. The demand segmentation, however, can be a difficult task, and it may even be possible that over some time intervals the stationary conditions are not satisfied. In addition, the dynamic features of the variables were not taken into account. Therefore, for managing these issues, a preprocessing procedure that transforms the data for achieving weakly stationary conditions and for dealing with the dynamic features of the network is used. PCA can then be applied for leak detection by using a statistic measure such as the combined index φ. The PCA-based leak location strategy allows estimating the contribution of each variable for identifying the potential zone where the leak occurs.

The main advantages of this approach are the following: (i) it only requires hydraulic data from the DMA operating under normal conditions; (ii) it allows the detection of leaks with small outflow magnitude compared with the total demand of the network; (iii) it allows identification of the leak zone location; (iv) generally, satisfactory results can be obtained by only using one pressure head sensor within each zone.

The structure of the paper is the following. In Section 2, the modeling of WDNs, the consumer demand and the water losses are presented. In Section 3, the PCA formulation for fault detection and identification together with the respective assumptions regarding the data are presented. The data processing for satisfying the PCA assumptions is described in Section 4. Thus, Section 5 depicts the leak detection and location with an unsupervised approach. Section 6 introduces the Hanoi network as the case study with the corresponding demand patterns. The results and discussion are presented in Section 7, and finally, in Section 8 conclusions and directions for future work are given.

2. Water distribution networks

2.1. WDN modeling. Two main physical laws for flow in pipe systems under steady conditions are the conservation of mass and the conservation of energy. Therefore, the model of a WDN with N nodes and $t \in \mathbb{Z}$ associated with the sampling time under normal operation conditions can be described mathematically by the following facts:

- The net inflow must be equal to the net outflow for any node $n \in N$ of the network:

$$\sum_{i=1}^{b_n} q_i(t) = d_n(t), \tag{1}$$

where b_n is the number of branches connected to the node n, $q_i(t)$ denotes the flow of the branch i and $d_n(t)$ is the respective demand.

- The sum of pressure heads around any loop of the network is equal to zero. Thus, a loop with G water sources and L water drops is modeled by

$$\sum_{g=1}^{G} h_g(t) + \sum_{l=1}^{L} h_l(t) = 0, \tag{2}$$

where pressure heads $h_g(t)$ and $h_l(t)$ are associated with sources and drops, respectively.

- The relation between flow $q(t)$ and pressure head $h(t)$ for any component of the network is modeled by

$$h(t) = \theta q^{\gamma}(t), \tag{3}$$

where the parameter θ depends on the specific component and the exponent γ could have a value close to 2 (Houghtalen *et al*., 2010).

2.2. Water losses. The water loss scenarios considered in this work are single unreported and abrupt faults: the possibility of one burst with a small to moderate outflow $f_n(t)$ located at any node n of the network. The leakage outflow caused by a pipe break depends on the node pressure head $h_n(t)$ and the pipe break size. A realistic approach for expressing this relationship is given by

$$f_n(t) = C_e h_n(t)^{\gamma}, \tag{4}$$

where the emitter coefficient C_e is associated with the pipe break size and $\gamma = 0.5$ (Rossman, 2000). This leakage outflow $f_n(t)$ has an effect similar to that of the demand in the mass balance equation for the node, i.e.,

$$\sum_{i=1}^{b_n} q_i(t) = d_n(t) + f_n(t). \tag{5}$$

Hence, some features of the demand must be considered for distinguishing between leaks and demand deviations. A demand model that describes some of these features is described below.

2.3. Demand model. Water distribution networks are demand-driven. In other words, the behavior of the flows and pressures is conditioned by the consumers' demand (Olsson, 2006). Moreover, as was previously shown, the features of the demand play an important role in the detection of the water losses. Therefore, the characterization of the demand at each node according to a model similar to the one presented by Zhou *et al*. (2002) is proposed.

The demand model $d_n(t)$ at each node n is formed by a periodic component for a specific time horizon (daily, weekly, monthly, yearly) together with an autoregressive component such that

$$d_n(t) = d_{n_\psi}(t) + d_{n_\xi}(t), \tag{6}$$

where $d_{n_\psi}(t)$ is a periodic function with period Γ of τ sampling times. It represents the mean consumption pattern for each node; $d_{n_\xi}(t)$ is an autoregressive stationary function that represents the consumption uncertainty around the periodic function which is modeled as

$$\begin{aligned} d_{n_\xi}(t) = \phi_0 + \phi_1 d_{n_\xi}(t-1) + \phi_2 d_{n_\xi}(t-2) \\ + \cdots + \phi_p d_{n_\xi}(t-p) + \epsilon_n(t), \end{aligned} \tag{7}$$

where $\epsilon_n(t)$ is a Gaussian process with zero mean and a time-dependent autocorrelation. Thus, the variability of $\epsilon_n(t)$ changes over the period Γ because of the demand uncertainty.

The demand model presented in (6) plays an important role in leak detection and location. First, the function $d_{n_\psi}(t)$ is mainly responsible for the nonstationary behavior of the network variables. The periodic feature of this term, however, allows formulating a transformation, presented in the next section, for obtaining weakly stationary variables. Second, the dynamic feature of the function $d_{n_\xi}(t)$ causes the network variables to also present a dynamic behavior. This fact motivates the use of methods for leak detection and location that consider this feature such as the unsupervised approach presented below.

3. Advanced PCA for FDI

The starting point of PCA is a data set of cardinality m which is formed by a measurement vector with flows and pressure head in the case of the WDN

$$x(t) = [q_1(t), q_2(t), \ldots, h_1(t), h_2(t), \ldots]^T \in \mathbb{R}^m. \tag{8}$$

The historical data set of hydraulic signals is then organized as a matrix formed by p observations of the vector $x(t)$, which can be represented as

$$X = \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(p) \end{bmatrix} = \begin{bmatrix} x_1(1)\, x_2(1) \ldots x_m(1) \\ x_1(2)\, x_2(2) \ldots x_m(2) \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ x_1(p)\, x_2(p) \ldots x_m(p) \end{bmatrix} \in \mathbb{R}^{p \times m}. \tag{9}$$

The goal of the classical PCA method is to find a linear transformation matrix $P \in \mathbb{R}^{m \times a}$ that projects each vector of variables $x(t)$ from X to a space where the process information in terms of variability is preserved: $\tilde{x}(t) = x(t)P \in \mathbb{R}^a$. If there is a certain level of the information redundancy among the variables, the dimension of the new space can be reduced such that $a < m$, and the transformed variables are called principal components. The columns of the transformation matrix P are a eigenvectors associated with the most significant eigenvalues of the correlation matrix R of the historical data matrix X.

The fault detection and identification task with the traditional PCA model assumes the linearity of the process and the following:

A1: Each variable $x_i(t)$ of matrix (9) has a Gaussian distribution function.

A2: Each variable $x_i(t)$ of matrix (9) is weakly stationary. Furthermore, because of the different magnitudes of the components of vector $x(t)$, each variable $x_i(t)$ is scaled to zero mean and unit variance.

A3: The vector $x(t)$ is not time-correlated.

3.1. Fault detection. Fault detection based on PCA is usually performed by using two distance measures: T^2 and the squared prediction error (SPE). They are calculated for an observation vector $x(t_{\text{new}})$ as

$$T^2_{\text{new}} = x(t_{\text{new}})^T P \Lambda^{-1} P^T x(t_{\text{new}}), \tag{10}$$

$$\text{SPE}_{\text{new}} = r^T r, \quad r = (I - PP^T)x(t_{\text{new}}), \tag{11}$$

where $I \in \mathbb{R}^{m \times m}$ is an identity matrix and Λ is a diagonal matrix with a principal diagonal that is formed by a eigenvalues of the correlation matrix in descending order. The thresholds T^2_α and SPE_α for each statistic (10) and (11) are adjusted for a confidence level α according to the desired sensitivity for the detector and a boundary false alarm rate, as explained by Chiang *et al*. (2001).

T^2 measures the deviations in the principal component directions, and SPE measures the deviations with respect to the PCA model. Thus, there is a complementary nature between these two measures which can be summarized in a combined index (Yue and Qin, 2001)

$$\varphi_{\text{new}} = \frac{\text{SPE}_{\text{new}}}{\text{SPE}_\alpha} + \frac{T^2_{\text{new}}}{T^2_\alpha} = x(t_{\text{new}})^T M x(t_{\text{new}}), \tag{12}$$

$$M = \frac{(I - PP^T)}{\text{SPE}_\alpha} + \frac{P\Lambda^{-1}P^T}{T^2_\alpha}. \tag{13}$$

The threshold for this statistic ζ_α is calculated with a confidence level α as explained by Yue and Qin (2001). The primary advantage of using the combined index measure is that it simplifies the fault detection while no fault information is lost. Therefore, the interpretation of the index behavior is easier for the system operator than when the two indexes are used together.

3.2. Fault identification. When a fault has been detected, the subsequent task of the system's operator is to resolve the possible cause or causes of the anomalous behavior. This diagnostic stage is challenging when many variables are related to the process. Thus, the fault identification goal is to inform the operators and engineers about which observation variables (symptoms) are most relevant for diagnosing the fault (Kościelny *et al*., 2017), thereby focusing attention on the subsystem(s) where it is most likely that the fault occurred (Chiang *et al*., 2001). In particular, when using the statistic φ_{new} for fault detection, Yue and Qin (2001) propose the use of a reconstruction-based contribution (RBC) method for fault identification; the method estimates how much each variable contributes to taking the statistic out of its threshold. In recent works, this approach has been used for the fault diagnosis of real applications with satisfactory results (Beghi *et al*., 2016).

The reconstruction of the combined index φ_{new} along a specific variable direction reduces the effect of this variable over the detection index (Alcala and Qin, 2009). Thus, the reconstruction of a variable is proportional to the contribution of that variable to the deviation of φ_{new}. Therefore, the variables that contribute more to the deviation of φ_{new} are supposed to be the variables mostly associated with the fault.

The reconstruction-based contribution of each variable i to the deviation of φ_{new} is

$$\text{RBC}_i^{\varphi} = \frac{(\xi_i^T M x(t_{\text{new}}))^2}{\xi_i^T M \xi_i}, \tag{14}$$

where $\xi_i \in \mathbb{R}^m$ is a canonical vector that represents the variable direction, e.g., the direction of the first variable is $\xi_i = [1\,0\,0\,\ldots]$. The derivation of this contribution index is briefly explained below.

Consider a fault observation $x(t)$ represented by

$$x(t) = \hat{x}(t) + \xi_i \overline{f_i}, \tag{15}$$

where $\hat{x}_l(t)$ is the fault-free observation, ξ_i is the fault direction and $\overline{f_i}$ is the fault magnitude which is unknown. For estimating the value of the fault-free observation along each variable direction, the following reconstructed vector is defined:

$$z(t)_i = x(t) - \xi_i f_i, \tag{16}$$

where f_i is the fault magnitude estimated along the direction of variable i. The value of the detection index φ of a reconstructed vector in a specific direction i is

$$\varphi(z(t)_i) = \|z(t)_i\|_M^2. \tag{17}$$

Thus, a reasonable approach for estimating $\overline{f_i}$ is by minimizing (17), which can be done by taking the first derivative of $\varphi(z(t)_i)$ with respect to f_i and making it equal to zero. This step yields

$$\hat{f}_i = (\xi_i^T M \xi_i)^{-1} \xi_i^T M x(t_{\text{new}}). \tag{18}$$

The computation of (18) can be performed if $\xi_i^T M \xi_i$ is nonzero (complete reconstructability condition). The reconstruction-based contribution of φ_{new} is then

$$\text{RBC}_i^{\varphi} = \|\xi_i \hat{f}_i\|_M^2. \tag{19}$$

This result is equal to (14). RBC differs from traditional contributions by a scaling factor which depends on the variable direction i. Furthermore, ξ_i does not have to be necessarily a vector; it can be a column-like matrix representing a multi-dimensional fault or multiple sensor faults (Alcala and Qin, 2009).

3.3. Performance evaluation for leak detection and location. Leak detection analysis is based on three performance indexes for evaluating the results: false alarm rate (FAR), fault detection rate (FDR) and fault detection delay (FDD). The FAR and FDR are defined in percentages as

$$\text{FAR} = \frac{\text{Fs}}{\text{Nd}} \times 100\%, \tag{20}$$

$$\text{FDR} = \frac{\text{Ts}}{\text{Fd}} \times 100\%, \tag{21}$$

where 'Fs' is the number of times that a fault is detected when the network is operating normally, 'Nd' is the number of normal samples, 'Ts' is the number of times that a fault is detected when there is actually a fault (leak) affecting the network operation and 'Fd' is the number of fault samples. The FDD is the time that it takes for detecting a fault. A fault is detected if the statistic threshold is exceeded for over two continuous samples or more.

Leak location analysis is performed once the fault has been detected in order to identify the zone of occurrence. The leak location performance is evaluated within the pattern recognition framework as the fault classification rate (FCR) which can be defined as

$$\text{FCR} = \frac{1}{\text{Cl}} \sum_{i=1}^{\text{Cl}} \text{fcr}_i, \tag{22}$$

$$\text{fcr}_i = \frac{\text{Cs}_i}{C_i} \times 100\%, \tag{23}$$

where 'Cl' is the number of classes, fcr_i is the fault classification rate of class i, C_i is the true number of observations which belong to class i and Cs_i is the number of observations identified as part of class i.

In the context of leak location analysis, the evaluation consists in determining the classification rate fcr_i for

all the leaks that can be considered within each zone. Therefore, the index $\text{fcr}_i$is redefined for the leak location evaluation as

$$\text{fcr}_i^* = \frac{1}{n_i}\sum_{j=1}^{n_i} \text{fcr}_{ij}, \tag{24}$$

$$\text{fcr}_{ij} = \frac{\text{Cs}_{ij}}{C_{ij}} \times 100\%, \tag{25}$$

where n_i is the number of nodes within the zone i, C_{ij} is the true number of observations which belong to the i when a leak at node j is emulated, and Cs_{ij} is the number of times that the fault data were identified as part of zone i when a leak affects the node j within the zone. The index fcr_{ij} then evaluates the performance of the method when a leak is emulated at node j in zone i.

4. Data processing and augmented space

As previously explained, the application of the traditional PCA method for fault detection and identification relies on the assumptions A1, A2 and A3. The hydraulic data from matrix X do not explicitly satisfy A2 and A3: the variables are not weakly stationary because of the periodic component of the demand, and the vector $x(t)$ is time-correlated because of the dynamic feature of the demand. Therefore, two processing stages are used here for satisfying the two assumptions above.

4.1. Periodic preprocessing. The demand represented by (6) is a periodically stationary process according to the definition of Papoulis (1991). As the WDN behavior is demand-driven, it is assumed here that the signals of flow and pressure head have a similar structure. Thus, each signal $x_i(t)$ is considered as a periodically stationary process which can be transformed into a weakly stationary process as described by the following fact.

Proposition 1. (Quinoñes-Grueiro *et al.*, 2017) *Let $x_i(t)$ be a periodically stationary process with the structure given by (6) and*

$$t' = \begin{cases} t, & 0 < t \leq \Gamma, \\ \\ \text{mod}(t, \Gamma), & t > \Gamma, \end{cases} \tag{26}$$

where $\text{mod}(t, \Gamma)$ is the remainder resulting from the division of t by Γ. The process

$$x_i^*(t') = x_i(t') - \mathcal{E}_{t'}\{x_i\} \tag{27}$$

is stationary, where the periodic expected value of x_i for t' can be estimated off-line by

$$\mathcal{E}_{t'}\{x_i\} \simeq \frac{1}{J+1}\sum_{j=0}^{J} x_i(t' + j\tau) \tag{28}$$

with large enough $J+1$ periods of the variable $x_i(t)$ and $J \in \mathbb{Z}$.

Proof. From the features of the variable $x_i(t')$, the set of equations for $J+1$ periods can be written

$$\begin{aligned} x_i(t') &= x_{i_\psi}(t') + x_{i_\xi}(t'), \\ x_i(t'+\tau) &= x_{i_\psi}(t'+\tau) + x_{i_\xi}(t'+\tau), \\ &\vdots \\ x_i(t'+J\tau) &= x_{i_\psi}(t'+J\tau) + x_{i_\xi}(t'+J\tau). \end{aligned}$$

Hence, adding these equations, we get

$$\begin{aligned} &\sum_{j=0}^{J} x_i(t'+j\tau) \\ &\quad = \sum_{j=0}^{J} \big(x_{i_\psi}(t'+j\tau) + x_{i_\xi}(t'+j\tau)\big) \\ &\quad = (J+1)x_{i_\psi}(t') + \sum_{j=0}^{J} x_{i_\xi}(t'+j\tau). \end{aligned}$$

By considering the time-varying mean of $x_i(t')$ given in (28) and the mean of $x_{i_\xi}(t')$ given by

$$\mathcal{E}_{t'}\{x_{i_\xi}\} \simeq \frac{1}{J+1}\sum_{j=0}^{J} x_{i_\xi}(t'+j\tau)$$

with J large enough, Eqn. (28) is equivalent to

$$\mathcal{E}_{t'}\{x_i\} \simeq x_{i_\psi}(t') + \mathcal{E}_{t'}\{x_{i_\xi}\}. \tag{29}$$

Thus, if the time-varying mean (29) for each t' is substituted in the transformation equation (27), it is reduced to

$$x_i^*(t') = x_{i_\xi}(t') - \mathcal{E}_{t'}\{x_{i_\xi}\}. \tag{30}$$

■

Note here that the time-varying periodic mean of $x_i(t')$ must be obtained *a priori* from historical data, as part of the training procedure for FDI tasks. In consequence, $x_i^*(t') = x_i^*(t'+j\tau)$ with $j \in \mathbb{Z}$ can also be considered as weakly stationary. Note here that the time-varying periodic standard deviation $\mathcal{E}_{t'}\{x_i^2\}$ of $x_i(t')$ can be straightforwardly calculated by following the same procedure such that the signal $x_i^*(t)$ is also standardized. Thus, the resulting data matrix is X^*. The data equivalent to $J \geq 30$ periods of the variable with the largest period should be available such that the central limit theorem can be assumed to be valid (Montgomery and Runger, 2014). Thus, the time-varying mean estimated can be considered a good approximation of the time-varying mean of each variable's population at t'.

The demand model plays an important role for the periodic preprocessing because it is assumed that the demands determine the features of the network variables. Therefore, investigating the periodicity of the hydraulic data before the application of the proposed methodology is recommended. There are methods available in literature that can be used for determining if a variable has a periodic behavior, and, in addition, they can be used for estimating the period. The proposal presented by Wang *et al*. (2006) can be used, for instance, to estimate the period of the hydraulic variables if the structure represented by (6) is assumed.

4.2. Augmented data space and dynamic PCA. Pipe networks are dynamic systems, i.e., there is a time-dependent behavior of the hydraulic signals. Regardless of the stationary conditions achieved by using the periodic transformation, the dynamic properties of the data should be taken into account for fault detection and location. Otherwise, many false alarms would be generated by small disturbances, and the detection of the faults that affect the dynamic behavior of the network would be delayed or even missed.

In the framework of unsupervised statistical tools for FDI such as principal component analysis, Ku *et al*. (1995) propose to form the following extended data matrix:

$$X_l^* = \begin{bmatrix} x_1^*(1) \dots x_1^*(1-l) \dots x_m^*(1) \dots x_m^*(1-l) \\ x_1^*(2) \dots x_1^*(2-l) \dots x_m^*(2) \dots x_m^*(2-l) \\ \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ x_1^*(N) \dots x_1^*(N-l) \dots x_m^*(N) \dots x_m^*(N-l) \end{bmatrix}, \tag{31}$$

where l is the number of lags considered such that $X_l^* \in \mathbb{R}^{N \times (lm)}$. The importance of this transformation is that it allows for considering time correlations between variables when developing a data-based model such as the PCA model. Thus, the observation vectors $x_l^*(t)$ are not time-correlated, and the assumption A3 is satisfied. The application of PCA on the augmented data space is called dynamic principal component analysis (D-PCA). Yet as another preprocessing stage was applied before, the integration of the two processing stages with PCA will be called periodic dynamic principal component analysis (P-DPCA). In addition, according to Chiang *et al*. (2001), the following amount of data X_{th} is required for estimating a reliable PCA model:

$$X_{th} = \frac{a + 2am + a^2}{2m}, \tag{32}$$

where a is the number of PCs and m the number of variables.

5. Leak detection and location with P-DPCA

Unsupervised FDI methods do not require fault data for process monitoring. In this sense, the PCA method and its variant P-DPCA previously presented only require data from the normal process operation, and both of them take into account the linear relationships among the variables. This last fact is important because even when the sensors are placed far from each other, the variables are related because of the time correlation induced by the demand patterns from period to period.

The leak detection and location is considered here as a fault detection and identification problem. The fault detection approach can be performed in two stages. First, an off-line detection stage encompasses the following steps:

- Determine the periodic expected value $\mathcal{E}_{t'}\{x_i\}$ and the periodic standard deviation $\sigma_{t'}\{x_i\}$ for each measured variable x_i and t' by using a sufficiently high number of J periods of hydraulic data from the normal process operation.
- Generate a weakly stationary and standardized data set X^* of transformed vectors $x^*(t)$.
- Estimate the proper lag number l from the data in X^*.
- Form the extended matrix $X_l^* \in \mathbb{R}^{N \times (lm)}$.
- Estimate the PCA parameters (P_l^*, Λ_l^* and ζ_α) according to the number of selected components and the confidence interval α.

Second, the on-line detection stage involves the following:

- Process the new vector $x(t)$ by using the periodic transformation and also standardize it.
- Form the extended vector for some value of l so that we have $x_l^*(t) = [x_1^*(t), \dots, x_1^*(t-l), \dots, x_m^*(t), \dots, x_m^*(t-l)]$.
- Estimate the value of the combined index φ_{new}.
- If the combined index exceeds the threshold ζ_α for two continuous samples, a fault is detected. Otherwise, restart the on-line stage.

Once a fault is detected, the goal of the fault identification method is to recognize the variables most related to the leak occurrence. Thus, if each variable is uniquely associated with a zone of the DMA, then by determining the variable with the major contribution index the leak zone is identified. When applying the RBC$^{\varphi}$ method for leak zone identification with the vector $x_l^*(t)$, the contribution of a variable and its lagged values

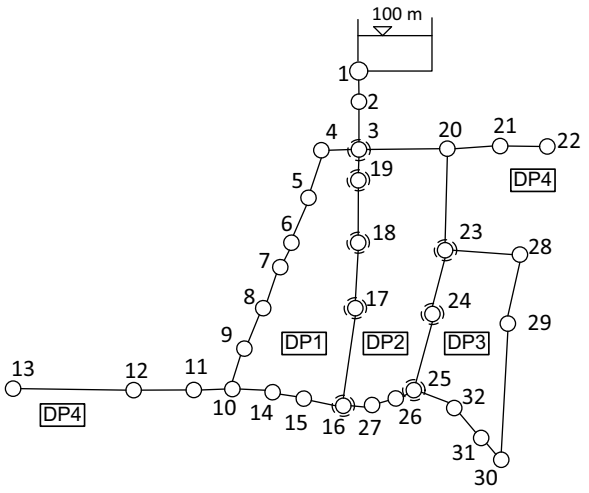


Fig. 1. Hanoi WDN.

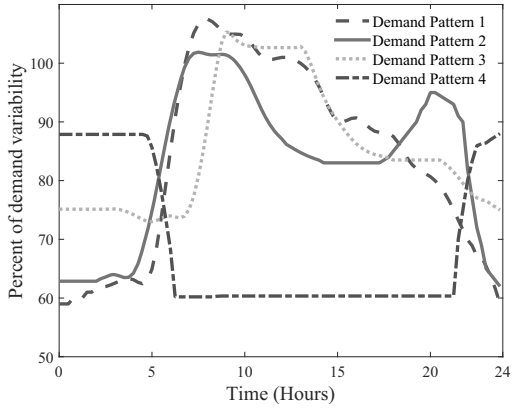


Fig. 2. Demand patterns $d_{i_\psi}(t)$ considered for the Hanoi WDN.

are associated with the same zone where the sensor is located. Hence, their RBC_i^φ are added for finding the total contribution of the specific zone.

6. Case study

The proposed approach was applied to the WDN from Hanoi, Vietnam, shown in Fig. 1. The network is formed by 32 nodes and 34 pipes distributed in three loops and two branches. No pumping equipment is required for this facility, and a single reservoir supplies water for the whole network. Many strategies have been applied for designing this network, which is fully described by Fujiwara and Khang (1990). In particular, the design (pipe diameters) proposed by Sedki and Ouazar (2012) is selected in this work. The four demand patterns (DPs) shown in Fig. 2 are considered for this network. The DP 1, 2 and 3 are of a residential type, and they correspond to the nodes in the loops of the network as presented in Fig. 1. The fourth DP is of a different type, and it corresponds to the nodes in the branches: nodes 11, 12, 13, 21 and 22. The nodes between two loops (surrounded by a dashed line) have an average demand pattern, e.g., nodes 23, 24 and 25.

6.1. Model simulator. The EPANET package (Rossman, 2000) together with MATLAB are used to obtain the simulated hydraulic data from the Hanoi network. The friction factor is calculated with the Hazen–Williams equation, and a roughness coefficient of 120 is used (Houghtalen *et al.*, 2010). The demand patterns $d_{i_\psi}(t)$ previously shown are simulated with daily periodicity ($\Gamma = 24$ hours). The sampling period considered for the simulations is 15 minutes, which gives $\tau = 96$ samples per period for each measured variable.

The consumption uncertainty is represented by

$$d_{i_\xi}(t) = 0.5 d_{i_\xi}(t-1) + \epsilon_i(t), \tag{33}$$

and the variance of $\epsilon_i(t)$ is assumed as a function of the periodic demand pattern given by $\sigma^2_{\epsilon_i}(t) = 0.03 d_{i_\psi}(t)$. Therefore, this time-varying variance allows emulating the impact of the consumption on the demand uncertainty, e.g., when there is more consumption, the demand uncertainty increases. The multiplying factor of 0.03 means that the uncertainty is considered here to be proportional to 3% of the total consumption of the respective node i.

7. Results and discussion

7.1. Sensor placement and zone division. In this work, it is assumed that the network is already designed by experience. The number of sensors used for monitoring the Hanoi network given its area extension and by considering a low-cost solution is supposed to be three. The location of measurement points within this network has not been optimized with respect to any performance criterion, but it has been selected according to the operators' experience: criteria such as the point with the highest elevation, the points at the edge of the network, and the points close to the big demands (Zhang *et al.*, 2016). Therefore, nodes 8, 20 and 31 are selected, i.e., x=$[h_8, h_{21}, h_{31}]$, because they are points associated with great demand nodes close to the edges of the network. Moreover, they cover pressure changes of different areas of the network.

The clustering of nodes per zone is a division problem. The zone division of WDN can be performed according to diverse criteria. For networks which are already instrumented, and without an explicit model available, the zone partition can be executed by using the experts' criteria as presented by Moczulski *et al.* (2016) and Zhang *et al.* (2016). If an explicit model is available, the k-means clustering algorithm can then be used such that the variables (nodes) with a similar pattern of variability are considered within the same zone (Zhang *et al.*, 2016).

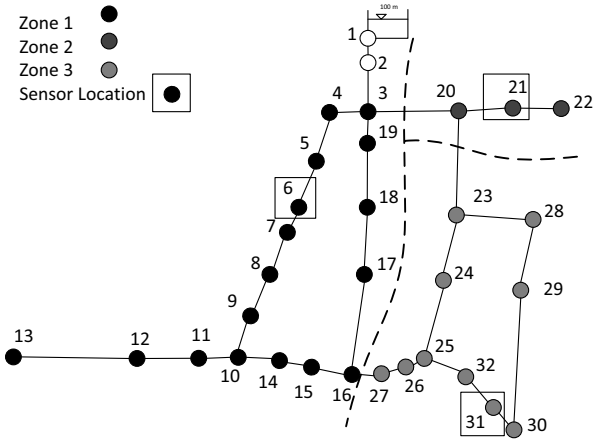


Fig. 3. Measurement points and zones of the Hanoi WDN.

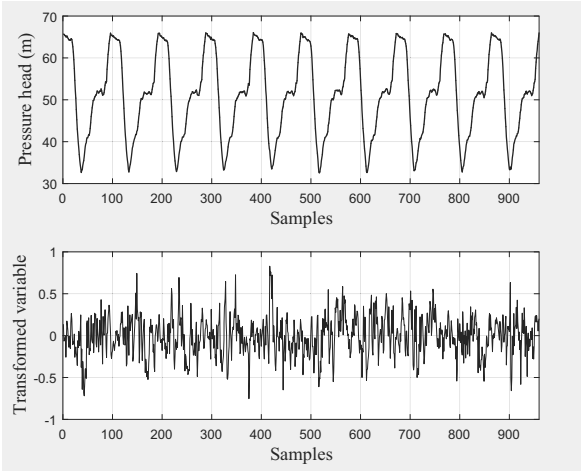


Fig. 4. Periodic transformation over a set of pressure head data from node 31.

A reasonable approach considered in this work is that the zone division task can be performed according to the demand pattern associated with each measurement sensor because the WDN are demand-driven systems such that the pressure at any component of the network will depend ultimately on the consumers' demand (Olsson, 2006). Therefore, the information regarding the pressure pattern per sensor (if available) can also be used for determining the nodes belonging to each zone. Given the number of sensors considered in this work, three leakage zones can be isolated. The main purpose of leak zone isolation is to reduce the search time for the operators given the enclosed area associated with each zone. Thus, the three measurement points and the zones of the network are shown in Fig. 3. Even when zone 2 covers only three nodes, the distance among these nodes is significant considering the length of the pipes (more than 1000 m).

Usually, the number of sensors available for installation in a WDN is small for economic reasons, and if the size of the network is large, then each zone will have a larger extension. There are two practical implications of these facts. First, the larger the network, the more sensors are required for guaranteeing the detection of leaks with a small outflow. Second, in case the network is not instrumented yet, the sensors should be placed such that the maximum isolation is achieved among the zones. The zone division and sensor placement within each zone for detecting all leaks and achieving maximum isolation at the same time is a subject which has not been yet addressed to the best of our knowledge. Therefore, it deserves more attention, but it is out of the scope of this work.

7.2. Method parameters. The Hanoi WDN was simulated under normal conditions with the given features of the demands for over 30 days to obtain the training historical data set. Thus, these pressure data are preprocessed according to (27) with $J = 30$. Figure 4 shows the effect of the transformation on pressure data from node 31 for over 10 periods (960 samples).

Method 1 proposed by Rato and Reis (2013) is used for the selection of the lag parameter. Therefore, $l = 3$ because by adding more lags new linear relationships do not seem relevant given that a small key singular value and the lowest key singular value ratio are obtained. After these steps, the data matrix (31) is formed. The information criterion is used for the selection of the number of principal components. The desired variability to be preserved is selected as 85%, and 4 principal components out of 12 are retained. The thresholds of the statistics were estimated by considering a confidence interval of $\alpha = 0.05$.

7.3. Leak scenarios. A leak scenario is characterized by the node location, the emitter coefficient, the time of appearance and the duration. A single leak was simulated for each node with an emitter coefficient value equal to 10 except for nodes 1 and 2. Node 1 is the source node, and the pressure head at node 2 is dominated by the pressure of the reservoir. Because the leak outflow size depends on the node pressure, it does not have a constant value, but it varies approximately between 18 [lps] and 40 [lps] (0.6% to 1.3% of the total demand). This represents less than 2.5% of the total demand of the network and between 3% and 25% of the node demand. All leaks appear at the zeroth hour, and each one lasts for 24 hours, which means that 96 samples are collected. No leak data was used to tune the parameters of the proposal.

7.4. Leak detection analysis. A total of 30 additional periods of normal pressure head data are simulated for testing the FAR. The average FAR obtained per period is 3%, and the behavior of the combined statistic for the 720

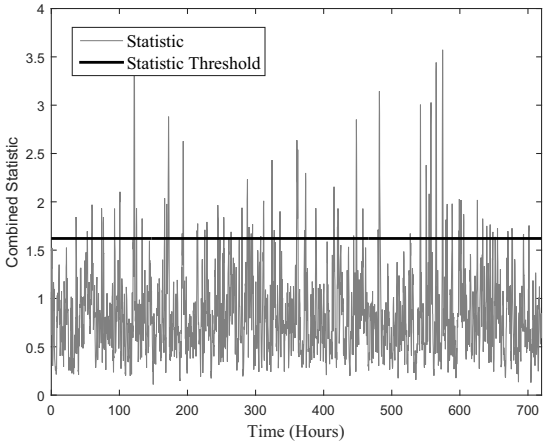


Fig. 5. Behavior of the combined statistic for 720 hours of normal data.

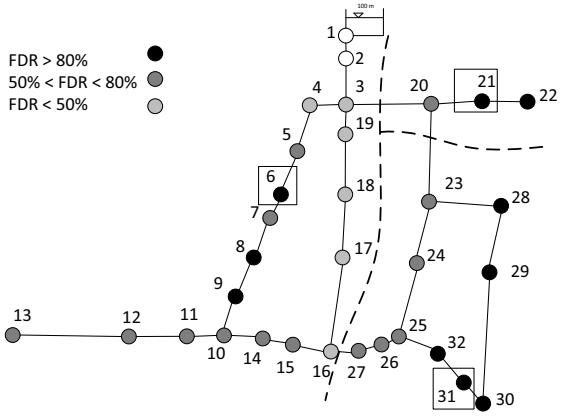


Fig. 6. FDR results for the Hanoi WDN with an emitter coefficient equal to 10.

hours is shown in Fig. 5. Note that from 2880 samples only 79 times a fault was detected.

The FDR results are represented in Fig. 6 for this network. In this work, less than 50% of FDR is considered unacceptable. Hence, the leaks with varying outflow rates between 18 [lps] and 40 [lps] are not detected satisfactorily by the proposal at nodes 3, 4, 16, 17, 18 and 19. On the one hand, the pressure head at nodes 3 and 4 is still dominated by the reservoir, and pressure changes are not evident. On the other hand, the other nodes are located on the border of the two zones defined for this network such that they are located too far from the measurement points. In addition, the FDRs obtained for the leaks in nodes 5, 7 and 20 are 58%, 64% and 50% even when they are located close to the measurement points. The reason for these results is that these nodes have a great demand which dominates the pressure head such that the change caused by a leak with the outflow simulated in this work is not too relevant.

The FDD index has an average value between 10 and 15 minutes except for the leaks at nodes 2 and 18 with 45 and 65 minutes, respectively. It is not strange that almost all the leaks are quickly detected because the time of appearance is in the early morning when the demand is relatively constant. If leaks were introduced at peak hours of consumption, however, the performance indexes would not be seriously affected as shown by Quiñones-Grueiro *et al*. (2017).

Even though the FDR results are not satisfactory for all leaks, note that they were detected within a time interval of 24 hours. The low FDR is caused by the uncertainty effect on the pressure such that because the high demand variability the leak effect is masked. This fact occurs especially throughout the high consumption time intervals and when the consumers' demand is changing drastically.

Furthermore, to test the proposed method for leaks with a greater outflow rate compared with the ones already simulated, the experiment was repeated with an emitter coefficient of 14. The resulting leaks have a varying outflow rate between 35 [lps] and 50 [lps] (1.1% to 1.6% of the total demand). The FDR results obtained are satisfactory for all the leaks except for the one in node 19. The demand in this node has a maximum outflow rate of 21.6 [lps] which is not significant compared with the closest nodes: 484.2 [lps] and 306 [lps] for nodes 18 and 3, respectively. This fact causes that the pressure head at node 19 is governed by the demand of nodes 3 and 18.

If the proposed approach is applied from 12 A.M. to 5 A.M., when the uncertainty is small, the results show that the average FDR index increases from 82.42% to 84.75% for an emitter coefficient equal to 14 (1.1% to 1.6% of the total demand). Despite the expected improvement, note that the proposal detects the leaks continuously. This fact is important given that if a leak with a significant outflow occurred during the day, it would remain undetected until the early morning such that it could represent an important loss of water with an economic cost and service deterioration.

7.5. Leak location analysis. The leak location performance results of the proposed method are presented in Fig. 7 for this network by using the index fcr_{ij} for each node. Unsatisfactory results are considered if fcr_{ij} has a value less than 50%. Hence, the leaks at nodes 16 and 17 cause the worst location results. It is remarked that the nodes with the worst classification results are located at the boundary among the zones. This is a reasonable result given that the effect of the leaks is similarly reflected in the measurement points of adjacent zones. Finally, the indexes fcr_i of the zones are 79.34%, 88.40% and 88.11%, respectively. The FCR for the proposed method is then equal to 85.28%.

The performance of the proposed unsupervised

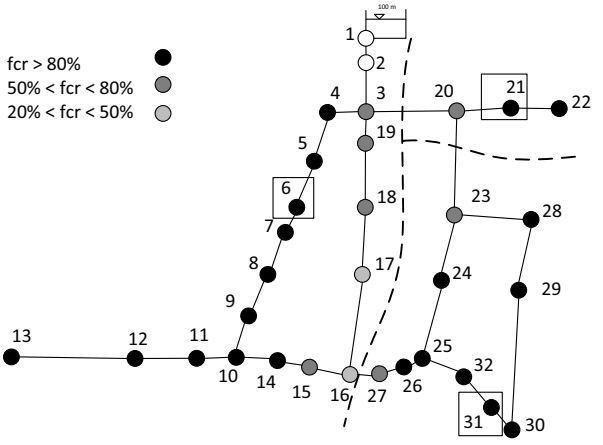


Fig. 7. Values of FCR for the Hanoi WDN with an emitter coefficient equal to 10.

approach is compared with the supervised one by assessing two classification methods: the Bayesian classifier and the K-nearest neighbor, reported by Soldevila *et al.* (2017; 2016) respectively. A day of data which consists of the residuals of the pressure head at each measurement point was generated for each leak scenario by considering the same conditions previously established. The residual is the difference between the value of the pressure head in normal conditions and under a leak scenario. The Bayesian classifier is calibrated by using the Gaussian probability density function, and the number of neighbors selected for the K-nearest neighbor (KNN) classifier is 3. These configurations are similar to the ones used in Soldevila *et al.* (2016; 2017). Since the time horizon analysis is not applied for the unsupervised approach given the importance of detecting the leaks as soon as possible, it will not be applied for the supervised methods in order to establish a fair comparison. For the evaluation of the classification accuracy a 10-fold cross-validation is performed, and the fault classification accuracy is estimated for each zone. The results presented in Table 1 show that for the conditions of the experiments the proposal achieves similar performance without requiring data of each leak scenario.

Table 1. Fault classification rate per zone.

	fcr Index		
Zone	Bayesian	KNN	Unsupervised approach
1	94.5	91	79.34
2	56.3	68.1	88.4
3	53.7	75.1	88.11
Total	68.16	78.06	85.28

8. Conclusions

An unsupervised approach to water loss detection and location in water distribution networks is presented. Fault detection is performed by combining two preprocessing methods with principal component analysis. A periodic transformation and a data vector extension method are applied in the first stage because they process the non-stationary and dynamic data from WDNs for building the PCA model. In the second stage, a combined statistic φ which measures the deviations from the data-based model is used for detecting leaks. The zone division strategy together with the reconstruction-based contribution index RBC_i^{φ} are employed for leak zone location. The main advantage of this proposal is that it allows leak detection and leak zone location by only using data from a reduced number of sensors when the network operates normally.

The feasibility of the proposed approach has been tested on the Hanoi WDN. The results show that by only using three pressure head sensors the leaks with an outflow rate between 18 [lps] and 40 [lps] (less than 2.5% of the total network demand) can be detected with an average detection rate of 72%. The leak location achieves a classification rate of 85.28%. Future research can focus on the application to real networks, the development of zone division strategies for WDNs and methods for sensor placement within each zone.

Acknowledgment

The authors wish to thank the editor and the anonymous reviewers for their thought-provoking and insightful comments and suggestions, which have been very useful for improving the paper quality and presentation. The authors also acknowledge the support provided by DGAPA-UNAM IT100716, II-UNAM and Universidad Tecnológica de La Habana José Antonio Echeverría (CUJAE).

References

Aksela, K., Aksela, M. and Vahala, R. (2009). Leakage detection in a real distribution network using a SOM, *Urban Water Journal* **6**(4): 279–289.

Alcala, C.F. and Qin, S.J. (2009). Reconstruction-based contribution for process monitoring, *Automatica* **45**(7): 1593–1600.

Beghi, A., Brignoli, R., Cecchinato, L., Menegazzo, G., Rampazzo, M. and Simmini, F. (2016). Data-driven fault detection and diagnosis for HVAC water chillers, *Control Engineering Practice* **53**: 79–91.

Chiang, L.H., Rusell, E. and Braatz, R.D. (2001). *Fault Detection and Diagnosis in Industrial Systems*, Springer, London.

Colombo, A.F. and Kamey, B.W. (2002). Energy and costs of leaky pipes: Toward comprehensive picture, *Journal of Water Resource Planning and Management* **128**(6): 441–450.

Fujiwara, O. and Khang, D.B. (1990). A two-phase decomposition method for optimal design of looped water distribution networks, *Water Resources Research* **26**(4): 539–549.

Houghtalen, R.J., Akan, A.O. and Hwang, N.H.C. (2010). *Fundamentals of Hydraulic Engineering Systems*, 4th Edn., Prentice Hall, Englewood Cliffs, NJ.

Jung, D. and Lansey, K. (2015). Water distribution system burst detection using a nonlinear Kalman filter, *Journal of Water Resources Planning and Management* **141**(5): 1–13.

Ku, W., Storer, R.H. and Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis, *Chemometrics and Intelligent Laboratory Systems* **30**: 179–196.

Łangowski, R. and Brdys, M.A. (2017). An interval estimator for chlorine monitoring in drinking water distribution systems under uncertain system dynamics, inputs and chlorine concentration measurement errors, *International Journal of Applied Mathematics and Computer Science* **27**(2): 309–322, DOI: 10.1515/amcs-2017-0022.

Laucelli, D., Romano, M., Savic, D. and Giustolisi, O. (2016). Detecting anomalies in water distribution networks using EPR modelling paradigm, *Journal of Hydroinformatics* **18**(3): 409–427.

Kościelny, J.M., Syfert, M., Rostek, K. and Sztyber, A. (2017). Fault isolability with different forms of the faults–symptoms relation, *International Journal of Applied Mathematics and Computer Science* **26**(4): 815–826, DOI: 10.1515/amcs-2016-0058.

Moczulski, W., Wycz, R., Ciupke, K., Przystałka, P., Tomasik, P. and Wachla, D. (2016). A methodology of leakage detection and location in water distribution networks—The case study, *Conference on Control and Fault Tolerant Systems SysTol, Barcelona, Spain*, pp. 331–336.

Montgomery, D.C. and Runger, G.C. (2014). *Applied Statistics and Probability for Engineers*, 6th Edn., Wiley, Hoboken, NJ.

Mounce, S.R., Mounce, R.B., Jackson, T., Austin, J. and Boxall, J.B. (2014). Pattern matching and associative artificial neural networks for water distribution system time series data analysis, *Journal of Hydroinformatics* **16**(3): 617–632.

Nowicki, A., Grochowski, M. and Duzinkiewicz, K. (2012). Data-driven models for fault detection using kernel PCA: A water distribution system case study, *International Journal of Applied Mathematics and Computer Science* **22**(4): 939–949, DOI: 10.2478/v10006-012-0070-1.

Olsson, G. (2006). Instrumentation, control and automation in the water industry—State-of-the-art and new challenges, *Water Science and Technology* **53**(4–5): 1–16.

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

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes*, 3rd Edn., McGraw-Hill, New York, NY.

Quiñones-Grueiro, M., Verde, C. and Llanes-Santiago, O. (2017). Features of demand patterns for leak detection in water distribution networks, *in* C. Verde and L. Torres (Eds.), *Modeling and Monitoring of Pipelines and Networks*, Springer, Cham, Chapter 9, pp. 171–189.

Quiñones-Grueiro, M., Verde, C. and Prieto-Moreno, A. (2016). Leaks' detection in water distribution networks with demand patterns, *3rd Conference on Control and Fault Tolerant Systems SysTol, Barcelona, Spain*, pp. 313–318.

Rato, T.J. and Reis, M.S. (2013). Defining the structure of DPCA models and its impact on process monitoring and prediction activities, *Chemometrics and Intelligent Laboratory Systems* **125**: 74–86.

Romano, M., Kapelan, Z. and Savić, D.A. (2010). Real-time leak detection in water distribution systems, *Water Distribution Systems Analysis Conference, ASCE, Tucson, AZ, USA*, pp. 1074–1082.

Romano, M., Kapelan, Z. and Savic, D.A. (2013). Geostatistical techniques for approximate location of pipe burst events in water distribution systems, *Journal of Hydroinformatics* **15**(3): 634–652.

Rossman, L.A. (2000). Epanet 2 User's Manual, *Technical report*, United States Envionmental Protection Agency, `http://www.epa.gov/nrmrl/wswrd/dw/epanet.html`.

Sanz, G., Pérez, R., Kapelan, Z., Savic, D. and Asce, A.M. (2015). Leak detection and localization through demand components calibration, *Journal of Water Resources Planning and Management* **142**(2): 1–13.

Sedki, A. and Ouazar, D. (2012). Hybrid particle swarm optimization and differential evolution for optimal design of water distribution systems, *Advanced Engineering Informatics* **26**(3): 582–591.

Soldevila, A., Blesa, J., Tornil-Sin, S., Duviella, E., Fernandez-Canti, R.M. and Puig, V. (2016). Leak localization in water distribution networks using a mixed model-based/data-driven approach, *Control Engineering Practice* **55**: 162–173.

Soldevila, A., Fernandez-Canti, R.M., Blesa, J., Tornil-Sin, S. and Puig, V. (2017). Leak localization in water distribution networks using Bayesian classifiers, *Journal of Process Control* **55**: 1–9.

Wachla, D., Przystalka, P.and Moczulski, W. (2015). A method of leakage location in water distribution networks using artificial-neuro fuzzy system, *IFAC-PapersOnLine* **48**(21): 1216–1223.

Wang, J., Member, S., Chen, T., Member, S. and Huang, B. (2006). Cyclo-period estimation for discrete-time cyclo-stationary signals, *IEEE Transactions on Signal Processing* **54**(1): 83–94.

Wu, Y., Liu, S., Wu, X., Liu, Y. and Guan, Y. (2016). Burst detection in district metering areas using a data driven clustering algorithm, *Water Research* **100**: 28–37.

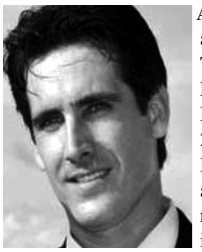
Yue, H.H. and Qin, S.J. (2001). Reconstruction-based fault identification using a combined index, *Industrial & Engineering Chemistry Research* **40**(20): 4403–4414.

Zhang, Q., Wu, Z.Y., Zhao, M., Qi, J., Huang, Y. and Zhao, H. (2016). Leakage zone identification in large-scale water distribution systems using multiclass support vector machines, *Journal of Water Resources Planning and Management* **142**(11): 1–15.

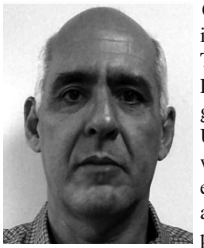
Zhou, S.L., McMahon, T.A., Walton, A. and Lewis, J. (2002). Forecasting operational demand for an urban water supply zone, *Journal of Hydrology* **259**(1–4): 189–202.

Marcos Quiñones-Grueiro received a BS in automation engineering in 2012, an MS in industrial computing and automation, and a PhD in technical sciences in 2017 from Universidad Tecnológica de la Habana José Antonio Echeverría (CUJAE). He is currently a full professor in the Automation and Computing Department at CUJAE. His research interests are in fault detection and diagnosis of industrial processes and water networks.

Cristina Verde received the BS and MS degrees in electronic communication and electrical engineering from National Polytechnic, Mexico, and the PhD degree in electrical engineering from Duisburg University in 1983. Prof. Verde then joined the National University in Mexico and received the 2005 Sor Juana Ines de la Cruz Award for outstanding women. She is an author of over 150 papers published at congresses and in journals. Her main topic of research is robust fault detection of pipelines and networks. She is a member of the IEEE, the Mexican National Academy of Engineering, and the Mexican National Academy of Sciences.



Alberto Prieto-Moreno graduated as an automation engineer from Universidad Tecnológica de la Habana José Antonio Echeverría (CUJAE) in 2004. He received his MSc in industrial informatics and automation in 2007, and his PhD in applied sciences in 2013. He is currently a full professor in the Automation and Computing Department of CUJAE. His research interests are in fault diagnosis of industrial systems, statistical pattern recognition, and artificial intelligence.



Orestes Llanes-Santiago received the BEng in electrical engineering from the Universidad Tecnológica de la Habana José Antonio Echeverría (CUJAE) in 1981. He pursued graduate studies from 1989 to 1994 at the University of the Andes in Mérida, Venezuela, where he obtained the MSc degree in control engineering in 1990 and the PhD degree in applied sciences in 1994. He is currently a full professor in the Automation and Computing Department at the Havana University of Technologies José Antonio Echeverría. His areas of interest are fault diagnosis, nonlinear control and computational intelligence.

Received: 24 March 2017
Revised: 6 September 2017
Re-revised: 11 January 2018
Accepted: 13 January 2018