Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 50-1 (2017) 3263–3268

Demand Model in Water Distribution Networks for Fault Detection [★]

Marcos Quiñones-Grueiro [*] Cristina Verde [**]
Orestes Llanes-Santiago [*]

[*] *Department of Automation and Computing, CUJAE, Habana, Cuba*
(e-mail: marcosqg@electrica.cujae.edu.cu, orestes@tesla.cujae.edu.cu).
[**] *Instituto de Ingeniería-UNAM Coyoacan 04510, México D.F.*
(e-mail: verde@unam.mx)

Abstract: A water distribution network (WDN) is a dynamic system in which the demand is a nonstationary process. On the other hand, most of the successful data-driven fault detection and isolation (FDI) methods have been developed by assuming static models and stationary processes. A demand model for the WDN is formed by a periodic signal plus a stochastic variable. This model allows the transformation of data such that a stationary and extended space of data can be obtained as it is demonstrated here. This proposition is the principal contribution of this work. To illustrate the advantages of the proposal, three well-known data-driven FDI algorithms are applied for the leak detection of the Hanoi distribution network: principal component analysis (PCA), independent component analysis (ICA), and support vector data description (SVDD). The leaks were emulated with different outflow magnitudes in all nodes. As a performance index, the fault detection rate is used, and the results indicate an improvement in the general index of approximately 75% when data are periodically transformed according to the proposal.

Keywords: Water distribution systems, Leak detection, Demand model, Periodic transformation, Data-driven methods for FDI.

1. INTRODUCTION

Water distribution networks (WDNs) are complex systems governed by main physical laws, system layout and consumers' demand and play an important role in urban life in meeting the demand for drinking water. Faults can produce water losses, issues with serious economic and ecological consequences and damage to the infrastructure (Pérez et al. (2014); Wu et al. (2016)). Hence, fault detection and isolation for WDNs is a topic of great interest. In particular, for the real scenario when only measurements of the pressure head and flow are available in some points of the network.

Model-based FDI approaches have achieved successful results for WDN monitoring (Pérez et al. (2014)). The identification of an analytical model, however, is not feasible for large real networks because of the amount of parameters, constraints and uncertainties involved. On the other hand, the network state obtained from instrumentation and supervisory control and data acquisition systems (SCADA) allows the use of data-driven FDI approaches.

Some data-driven FDI methods have been recently designed for WDNs. Artificial neural networks, principal component analysis, support vector machines, and clustering methods have been proposed (Mounce et al. (2011); Arsene et al. (2012); Arregui and Carlos (2012); Wu et al. (2016)). These works assumed the following: 1) flow sensors are installed in some nodes or the pressure head is measured in most of the nodes; 2) uncertainty is only related to the measurements; and 3) the demand variability is neglected during the day. In addition, only outflows of the reservoirs and the pressure head at some points of the network are measured. The water demand is a nonstationary process, and dynamic effects are present. Most data-driven FDI approaches can only be applied then in the early morning when the process is almost stationary with low variability.

Zhou et al. (2002) proposed a demand model in a network formed by additive periodic components and an autoregressive component for prediction tasks. The consideration of a similar demand model allows formulating the fault detection scheme proposed here. As a consequence, by assuming the pressure head of the reservoir as constant and the system linearity, the network variables can be described as well with periodic patterns in normal operation. A primary advantage of this consideration is that the data set can become stationary following the idea suggested by Wu et al. (2016) that the measurements of a WDN at a specific time do not vary much from day to day. In the latter work, the leak detection is formulated as an outlier detection problem, but it does not explicitly consider that each variable in the network is correlated with itself because the whole system is dynamic.

Thus, the main contributions of this work are as follows: 1) a demand model formed by a known periodic pattern plus a stochastic process; 2) the time variant transformation of a periodically stationary variable such that it becomes

[★] Paper supported by DGAPA-UNAM IT100716, II-UNAM and Universidad Tecnológica de La Habana José Antonio Echeverría (CUJAE).

2405-8963 © 2017, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2017.08.460

stationary; and 3) the generation of the extended space of data for dynamic systems which allows the application of most data-driven FDI methods. The advantages of the integrated proposal are discussed with three data-driven FDI methods for the Hanoi WDN, where leaks of different outflow magnitudes are emulated. The leak detection rate shows a considerable improvement of around 75% with the three methods if the data are previously transformed according to the proposal. Moreover, the on-line technical implementation requirements can easily be fulfilled since only vectors, transformation matrices and variables must be stored, and on-line math operations are not complex, e.g. multiplication and addition.

The content of this work is the following. In Section 2 the WDN with the demand model is explained. This section includes a straightforward transformation for obtaining a stationary process from a periodically stationary process and the extended data matrix which captures the data dynamics for a WDN. Section 3 briefly describes the three data-driven FDI methods which are considered for analyzing the effect of the data processing proposal. Section 4 introduces the case study with the demand patterns. Results and discussion are given in Section 5, and finally Section 6 presents the conclusions and future direction for the application of the proposal with real data.

2. DEMAND MODEL AND DATA PROCESSING

A WDN can be briefly described as a system which provides water according to the user demand and is formed by reservoirs, interconnected pipelines at some nodes and a set of demands points (Houghtalen et al. (2010)). The behavior of the flow in the pipes and the pressure head at the nodes is governed by physical laws, the system layout and the consumer demand at some nodes. Here the faults are considered leaks at the nodes. In general, a set of measured variables given by the following is assumed:

$$x(t) = [h_1(t), h_2(t), ..., h_i(t), q_1(t), q_2(t), ..., q_j(t)]^T \in \Re^m$$

where $h_i(t)$and $q_j(t)$ represent the i pressure head and the j flow of the network respectively and $t \in \mathbb{Z}$ is associated with the sampling time.

Because the consumer demand is an uncertain process and affects the networks' operation in a similar way to that of a leak at a node, to distinguish a demand variation from a leak some features must be assumed for the demand pattern independently of the used sensors. Zhou et al. (2002) propose a demand model for forecasting purposes that consists of three additive components. Their study indicates, however, that the demand variation is characterized by periodic patterns with a slowly varying component and added random variations. This result allows modeling the demand by two additive components: one that describes the periodic behavior and another that accounts for the additional variability. Thus, the demand for each node i is considered in this work as a periodically stationary process modeled as

$$d_i(t) = d_{i_\psi}(t) + d_{i_\xi}(t) \tag{1}$$

where $d_{i_\psi}(t)$ is a periodic function with period Γ of τ sampling times, which describes the behavior of the outflow at node i, and $d_{i_\xi}(t)$ is associated with the uncertainty around the periodic function. This term is considered as an autoregressive stationary process

$$d_{i_\xi}(t) = \phi_0 + \phi_1 d_{i_\xi}(t-1) + ... + \phi_p d_{i_\xi}(t-p) + \epsilon_i(t) \tag{2}$$

with $\epsilon_i(t)$ a weakly stationary process with a constant expected value and time-dependent autocorrelation, because its variability changes over the period Γ (Papoulis, 1991). While the periodic term must not have a serious impact on the results, the uncertain term determines the performance such that if the uncertainty increases throughout a specific time interval the performance will decrease given that the leaks can smear into the normal data. An advantage of the features given to model (1) is that by using linearity considerations the elements of $x(t)$ are also periodically stationary processes with period Γ. Therefore, if the periodicity of the vector $x(t)$ could be canceled out, all the data-driven FDI methods which require weakly stationary variables can be directly applied. The following subsection introduces a periodic transformation which cancels the periodicity of $x(t)$ and retains its variability.

2.1 Data Preprocessing

A periodically stationary process $x_i(t)$ with the structure given by (1) can be transformed into a weakly stationary process by using its periodic expected value at each t as it is described in the following fact.

Fact: Let $x_i(t)$ be a periodically stationary process with the structure given by (1) and

$$t' = \begin{cases} t & 0 < t \leq \Gamma \\ mod(t, \Gamma) & t > \Gamma \end{cases} \tag{3}$$

where $mod(t, \Gamma)$ is the remainder of the division of t by Γ, then the process

$$x_i^*(t') = x_i(t') - \mathcal{E}_{t'}\{x_i\} \tag{4}$$

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

$$\mathcal{E}_{t'}\{x_i\} \simeq \frac{1}{J+1} \sum_{j=0}^{J} x_i(t' + j\tau) \tag{5}$$

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

Since the proposed time-variant transformation depends on historical data of the system under normal conditions, the time-variant periodic mean of $x_i(t')$ must be obtained a priori as part of the training procedure for FDI tasks. As a consequence of this fact, $x_i^*(t') = x_i^*(t' + j\tau)$ with $j \in \mathbb{Z}$ can also be considered as weakly stationary.

The fact can be straightforwardly proven as follows.

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

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

Hence, by adding the left and right sides of these equations, the relation

$$\sum_{j=0}^{J} x_i(t'+j\tau) = \sum_{j=0}^{J} \big(x_{i_\psi}(t'+j\tau) + x_{i_\xi}(t'+j\tau)\big)$$
$$= (J+1)x_{i_\psi}(t') + \sum_{j=0}^{J} x_{i_\xi}(t'+j\tau) \quad (6)$$

can be written. By considering the time-variant mean of $x_i(t')$ given in (5) 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 a large enough value, Eq. (5) is equivalent to

$$\mathcal{E}_{t'}\{x_i\} \simeq x_{i_\psi}(t') + \mathcal{E}_{t'}\{x_{i_\xi}\} \quad (7)$$

Thus, if the time-variant mean (7) for each t' is substituted in the transformation equation (4), this is reduced to

$$x_i^*(t') = x_{i_\xi}(t') - \mathcal{E}_{t'}\{x_{i_\xi}\} \quad (8)$$

Therefore, since $x_{i_\xi}(t')$ is weakly stationary, the variable $x_i^*(t')$ is as well a weakly stationary variable.

A property of the transformation of each $x_i(t)$ according to (4) is that the time correlation of the variable is held.

2.2 Data Vector Extension

The system description with variables $x(t)$ in general via the data matrix X^*

$$X^* = \begin{bmatrix} x_1^*(1) & x_2^*(1) & \cdots & x_m^*(1) \\ x_1^*(2) & x_2^*(2) & \cdots & x_m^*(2) \\ \vdots & \vdots & \cdots & \vdots \\ x_1^*(N) & x_2^*(N) & \cdots & x_m^*(N) \end{bmatrix} \quad (9)$$

with $N = J\tau$ samples does not reveal the exact relations between variables in the model if rows of X^* are time-correlated. These time correlations generate excessive false alarms, specially for small disturbances if data-driven FDI methods use X^*. Therefore, in accordance with Ku et al. (1995), if variables are time-correlated, the extended data matrix $X_l^* \in \Re^{N\times(lm)}$ must be formed by

$$X_l^* = \begin{bmatrix} x_1^*(1) & \ldots & x_1^*(1-l) & \ldots & x_m^*(1) & \ldots & x_m^*(1-l) \\ x_1^*(2) & \ldots & x_1^*(2-l) & \ldots & x_m^*(2) & \ldots & x_m^*(2-l) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_1^*(N) & \ldots & x_1^*(N-l) & \ldots & x_m^*(N) & \ldots & x_m^*(N-l) \end{bmatrix} \quad (10)$$

where l is the number of lags and can be selected as proposed by Rato and Reis (2013).

For the WDN, once the vector $x(t)$ is preprocessed by using (4) and the data matrix is properly extended according to (10), methods like PCA, ICA, and SVDD could be applied for FDI tasks.

3. DATA-DRIVEN FDI METHODS

The data-driven methods can be divided into supervised and unsupervised based on whether historical fault data are required or not. Unsupervised methods make leak detection feasible for real applications because while plenty of normal operation data is generally available, fault data is not abundant. Therefore, three popular data-driven FDI methods are selected to show the advantages of the proposed data processing. The procedure for the application of the proposed approach consists then of two parts: the training and validation which characterize off-line the specific WDN to be monitored under normal conditions and the detection part which estimates on-line the behavior of the network and determines abnormal conditions. Thus, in particular one has the following steps.

- Off-line: 1) Determine for each measured variable x_i and t' the expected value $\mathcal{E}_{t'}\{x_i\}$ by using a high enough number of J periods of the network in normal conditions. 2) Generate a set of transformed vectors with cardinality N by applying (4) in normal conditions. 3) Estimate the proper lag number l from the data in X^*. 4) Form the extended matrix $X_l^* \in \Re^{N\times(lm)}$ 4) Generate the specific FDI statistics and parameters for each of the three considered methods.
- On-line: 1) Preprocess the new vector $x(t)$ by using (4). 2) Form the extended vector $x_l^*(t)$ with l previous samples given by
$$[\, x_1^*(t) \;\ldots\; x_1^*(t-l) \;\ldots\; x_m^*(t) \;\ldots\; x_m^*(t-l)\,]$$
3) Estimate the respective FDI indices.

To distinguish the standard acronyms of the FDI methods with those modified with the preprocessing and the dynamic space, the capital letters P and D have been placed before the acronyms respectively.

3.1 Principal Component Analysis (P-D-PCA)

Principal component analysis is a feature extraction method which reduces the data space preserving most of the process' information in terms of variability (Chiang et al. (2001)). Given the covariance matrix of the processed data $X_l^* \in \Re^{N\times(lm)}$

$$\Sigma_l^* = \frac{X_l^{*\,\prime} X_l^*}{N-1} \quad (11)$$

the eigenvectors associated with the a most significant eigenvalues of Σ_l^* define the matrix $P_l^* \in R^{(lm)\times a}$ from which the row vector of uncorrelated signals $\tilde{x}_{l\,a}^*(t)^T \in \Re^a$ with $a \leq m$, called principal components (PCs), is directly obtained for each vector $x_l^*(t)^T \in \Re^{(lm)}$ by using $\tilde{x}_{l\,a}^*(t) = x_l^*(t)P_l^*$. In this work, the number of PCs retained (a) is selected according to the SCREE test (Chiang et al. (2001)).

Fault detection is performed by using two statistics to measure the signal deviation in the principal component subspace T^2 and the squared prediction error spanned by the reduced model Q. If each statistic value is smaller than its respective threshold, T_α^2 and Q_α, the system holds the normal condition. These are adjusted by considering the sensitivity of the detector and a boundary false alarm rate, as explained in Chiang et al. (2001).

3.2 Independent Component Analysis (P-D-ICA)

Independent component analysis is a feature extraction method which reduces the data representation by maximizing the statistical independence of the data (Hyvärinen et al. (2001)). ICA algorithms find the vector of signals $\tilde{x}_{l\,a}^*(t)^T \in \Re^a$ with $a < (lm)$, called independent components (ICs), that maximize a non-Gaussianity measure by using $\tilde{x^*_{l}}_a(t) = x^*_l(t)W^*_l$ with $W^*_l \in R^{(lm)\times a}$. The FastICA algorithm is used here for finding the ICs given its simplicity and wide application (Hyvärinen et al. (2001)).

Independent components are ordered in a similar way as PCA information criterion, which is based on the norm L_2 of the rows of W^*_l. A graphical technique to determine the number of ICs analogous to the SCREE test is then applied. Fault detection is performed by using three statistics (I, I_e and Q) to measure the signal deviation spanned in the independent component subspace (I, I_e) and the squared prediction error spanned by the reduced model (Q). Their respective thresholds are calculated with the kernel density estimation method.

3.3 Support Vector Data Description (P-D-SVDD)

Support vector data description is a method used for one-class classification problems based on the hypersphere in the feature space F with minimum volume that holds all data X^*_l (Tax and Duin (2004)). To achieve this, the following optimization problem is solved

$$\min_{R,b,\xi}\{R^2+C\sum_{t=1}^{N}\xi(t)\} \text{ such that } \parallel \Phi(x^*_l(t)^T)-b \parallel \leq R^2+\xi(k) \quad (12)$$

where b and R are the center and the radius vectors of the hypersphere, respectively, C denotes the trade-off between the volume of the hypersphere and number of errors, Φ represents the Gaussian kernel function with bandwidth h which projects the data into F, and ξ indicates the slack terms which represent the probability that some of the training samples are erroneously classified (Tax and Duin (2004)). A fault is detected if the distance between the projection of the new sample $\Phi(x^*_l(t))$ and the center of the hypersphere b is larger than the radius R.

4. CASE STUDY

The case study is a benchmark for network design presented by Fujiwara and Khang (1990), which describes the planned water distribution trunk network in Hanoi, Vietnam. The configuration of the network is depicted in Fig. 1. No pumping facilities are required. It is formed by 32 nodes and 34 pipes organized in three loops and two branches. The system is gravity-fed by a single reservoir, the design parameters (pipe diameters) were obtained from Sedki and Ouazar (2012) and the link lengths can be found in Fujiwara and Khang (1990). It is assumed that pressure head sensors are located at nodes $3, 10, 16, 23, 25$, and x=[$h_3, h_{10}, h_{16}, h_{23}, h_{25}$]. These nodes are selected considering the maximal connectivity of the branches.

4.1 Model Simulator

The package EPANET (Rossman (2000)) and MATLAB are used to simulate this network. The Hazen-Williams equation is utilized for the calculation of the friction factor with a roughness coefficient of 120 (Houghtalen et al. (2010)). The three periodic functions $d_{i_\psi}(t)$ used to obtain synthetic data are shown in Fig. 2. They are simulated with daily periodicity ($\Gamma = 24$ hours) and a sampling period of 5 minutes such that there are $\tau = 288$ samples per period for all nodes in each zone with their respective branches. For the pipelines located between two zones, the demand is averaged. The uncertain demand $d_{i_\xi}(t)$ at each node is modeled by

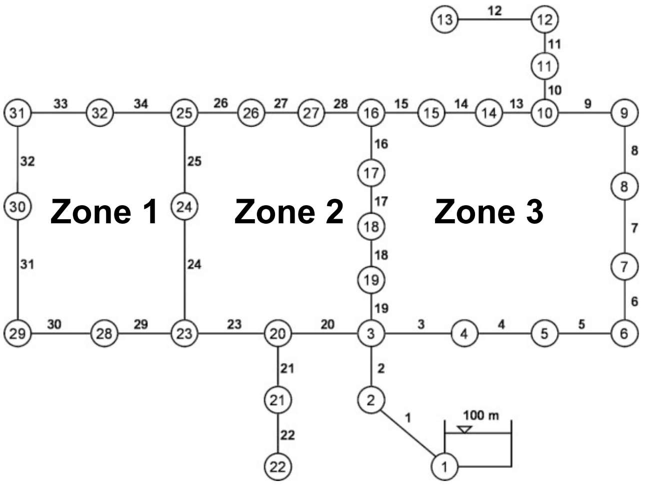

Fig. 1. Hanoi water distribution network

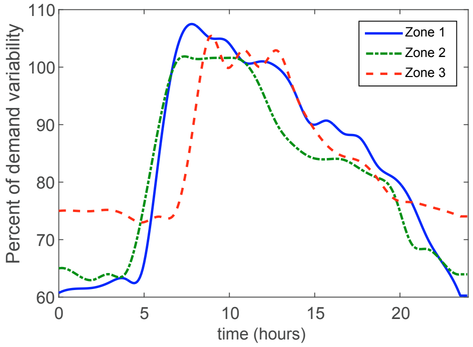

Fig. 2. Demand functions $d_{i_\psi}(t)$ for all nodes i in each zone

$$d_{i_\xi}(t) = 0.5d_{i_\xi}(t-1) + \epsilon_i(t), \quad (13)$$

and the variance of $\epsilon_i(t)$ is assumed as a function of the periodic demand pattern given by $\sigma^2_{\epsilon_i}(t) = 0.03d_{i_\psi}(t)$. For all simulated scenarios, white noise is added to the measured variables, and only a single leak occurs. A realistic approach is used to set an emitter coefficient (EC) C_e in a node that will generate a leakage outflow of magnitude f_i given by

$$f_i(t) = C_e h_i(t)^\gamma \quad (14)$$

where $h_i(t)$ is the pressure head at the node i and $\gamma = 0.5$ (Rossman (2000)).

5. RESULTS AND DISCUSSION

5.1 Model Parameters

Training data from the network under normal conditions were obtained with the given features for the demands over 30 weekdays. Thus, the preprocessing (4) with $J = 30$ is applied to the data. The lag parameter $l = 3$ was obtained to form the extended data matrix (10).

To retain 99% of the variability, 10 principal components were chosen with the P-D-PCA method. For the P-D-ICA method, 10 independent components were also selected. In the case of the P-D-SVDD method, the parameters b and R were obtained with $h = 5.3$ and $C = 0.27$. All

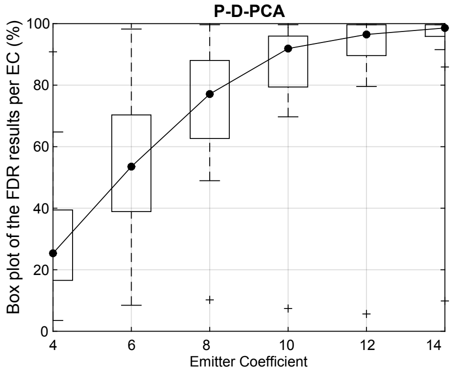

Fig. 3. FDR results of P-D-PCA for all the nodes with different emitter coefficients

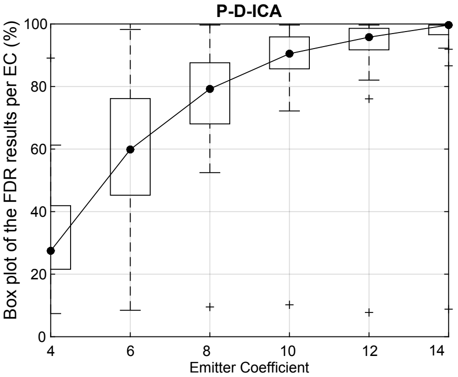

Fig. 4. FDR results of P-D-ICA for all the nodes with different emitter coefficients

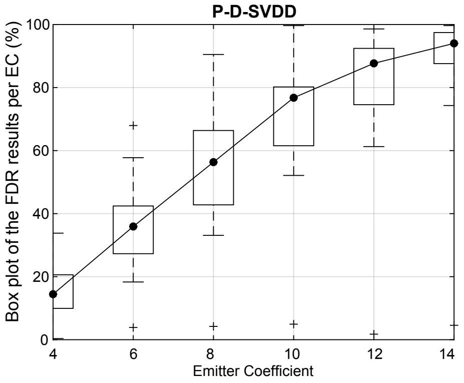

Fig. 5. FDR results of P-D-SVDD for all the nodes with different emitter coefficients

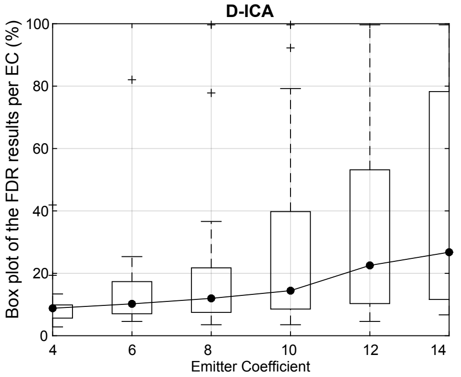

Fig. 6. FDR results of D-ICA for all the nodes with different emitter coefficients

thresholds of the statistics were estimated by considering a confidence interval of 0.05, and they were further adjusted to guarantee less than 5% of false alarm rate.

5.2 Leak Detection Scenarios

A single leak was simulated in each node of the network with EC values of 4, 6, 8, 10, 12 and 14 to test the sensitivity of the FDI methods such that the outflow rate depends on the node pressure representing mostly between 3% and 25% of the node demand. Each simulation scenario is then characterized by the node location of the leak and its respective emitter coefficient. All leaks were simulated over 288 samples.

The performance of the techniques for each scenario was evaluated by using the fault detection rate (FDR) index, which is defined as the number of times that the leak is detected divided by the total number of samples where there is a leak present. This rate permits developing a sensitivity analysis with respect to the leak outflow magnitude. The information of the graphics is compacted, and box plots are drawn to easily visualize the behavior of the index for all the scenarios. The symbol $\odot$ denotes the median, the height of the box represents the 75% of the index distribution, the dashed line $(--)$ expresses the minimum and maximum rate, and the symbol + describes the outlier of the distribution.

The results obtained by using the data processing under the same leak scenarios are presented in Figs. 3, 4 and 5 for P-D-PCA, P-D-ICA and P-D-SVDD, respectively. From these graphs, it is possible to conclude that all leaks could be detected within a 24-hour period even for small magnitude leak (less than 5% of the node demand).

The performance of the three tested algorithms together with the preprocessing depends on the leak outflow and the network uncertainties. If less than 50% of the median FDR index is considered as unacceptable, the results are then satisfactory for an EC greater than 6 (P-D-PCA, P-D-ICA) or 8 (P-D-SVDD). The evolution of the median FDR index for the three methods is similar such that when the leak size grows the median FDR index gradually increases, and the 75% of the index distribution decreases. The network uncertainties in these experiments are associated with the demand, that is, the nature of the term $\epsilon_i(t)$ (normal and weakly stationary) and the dynamic coefficients, and the measurement noise. Thus, for the same uncertainty conditions the performance of three algorithms improves for leaks with greater size. For real data, however, the nature of the term $\epsilon_i(t)$ must be investigated such that at least the weakly stationary conditions can be fulfilled. In addition, the instrumentation used in real applications can deteriorate the final performance, because of the sensors' low sensitivity.

The use of the two data processing stages is important for achieving these results, especially the transformation for achieving stationary signals from the network. Therefore, for comparison purposes, the D-ICA method is applied and results are shown in Fig. 6. The performance is significantly smaller compared to the one obtained with the proposed processing in Fig. 4. The median FDR index does not increase above 50% even for an EC of 14 (leak size between 36 lps and 50 lps). This illustrates for this case study that if the first processing method is not applied, although the dynamic behavior has been taken into account, there is a significant performance degradation.

Table 1 summarizes the median and standard deviation (μ, σ) of the FDR for the four methods with respect to the EC. The performance degradation without preprocessing is given by D-ICA in the last row. The Friedman and Wilcoxon hypothesis tests were used to evaluate the FDR of the methods (Demsar (2006)).

Table 1. Median and standard deviation (μ, σ) of the FDR for each emitter coefficient

Technique/C_e	4	6	8	10	12	14
P-D-PCA	(25, 18.90)	(53, 21.29)	(77, 19.13)	(91, 17.13)	(96, 17.20)	(98, 16.05)
P-D-ICA	(**27**, 18.50)	(**60**, 19.14)	(**80**, 18.05)	(90, 16.19)	(95, 16.63)	(**99**, 16.26)
P-D-SVDD	(14, 7.62)	(36, 13.34)	(56, 17.12)	(76, 16.80)	(87, 18.30)	(94, 17.45)
D-ICA	(09, 6.90)	(10, 14.03)	(11, 21.02)	(14, 26.80)	(22, 31.45)	(26, 35.31)

Even though the three methods show a similar index behavior, according to the hypothesis tests' results, the P-D-ICA method provides better detection rates for leaks with ECs 4, 6, 8, and 14 in this network. The performance of each of these three methods, however, depends on the nature of the raw data, the relationships among the variables involved and the parameter setting. Therefore, while there are straightforward procedures for the parameter setting of PCA and ICA for fault detection, the adjustment of the SVDD parameters is complicated. Thus, the use of SVDD for fault detection deserves more attention. Furthermore, even when there is a performance difference, it cannot be considered that the ICA variant is the best method among the three, but simply for this particular case study and simulation conditions.

6. CONCLUSIONS

To cope with the leak detection problem in a water distribution network, a demand model is proposed which includes dynamic and periodically stationary terms. It has been shown that an advantage of this model is that it allows the determination of a time-variant transformation for each network measured variable such that a stationary and extended data space are reached. As a consequence, this data preprocessing allows the application of data-driven FDI methods even for nonstationary time-correlated variables. To show the performance of the proposal for leak detection, three data-driven methods (PCA, ICA, and SVDD) are applied to the Hanoi network. The reported results make evident that even leaks with small outflows are detected for the case study. The fault detection rate shows an improvement of approximately 75% when the data are periodically transformed by the proposal. Future research can be focused on the extension of this proposal for fault location.

ACKNOWLEDGEMENTS

The authors acknowledge the support provided by DGAPA-UNAM IT100716, II-UNAM and the Universidad Tecnológica de La Habana José Antonio Echeverría (CUJAE).

REFERENCES

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.

Arsene, C.T.C., Gabrys, B., and Al-dabass, D. (2012). Decision support system for water distribution systems based on neural networks and graphs theory for leakage detection. *Expert Systems with Applications*, 39(18), 13214–13224.

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

Demsar, J. (2006). Statistical Comparison of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7, 1–30.

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.* Prentice Hall, 4th ed. edition.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis.* John Wiley & Sons, Inc.

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.

Mounce, S.R., Mounce, R.B., and 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.

Papoulis, A. (1991). *Probability, Random variables, and Stochastic Processes.* McGraw-Hill, 3rd edition.

Pérez, R., Sanz, G., Puig, V., Quevedo, J., Cugueró-Escofet, M.A., Nejjari, F., Meseguer, J., Cembrano, G., Mirats Tur, J.M., and Sarrate, R. (2014). Leak Localization in Water Networks. *IEEE Control Systems Magazine*, 34(4), 24–36.

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.

Rossman, L.A. (2000). *Water supply and water resources division. National Risk Management Research Laboratory. Epanet 2 User's Manual. Tech. Rep.* United States Envionmental Protection Agency. URL `http://www.epa.gov/nrmrl/wswrd/dw/epanet.html`.

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.

Tax, D.M. and Duin, R.P. (2004). Support Vector Data Description. *Machine Learning*, 54(1), 45–66.

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.

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.