Leak Localization in Water Distribution Networks using Pressure Models and Classifiers

Adrià Soldevila, Sebastian Tornil-Sin, Joaquim Blesa, Rosa M. Fernandez-Canti, and Vicenç Puig

Abstract This chapter proposes a leak localization architecture and an associated methodology in Water Distribution Networks (WDNs) using pressure models and classifiers. In a first stage of the proposed architecture, residuals are obtained by comparing available pressure measurements with the estimations provided by a WDN model. In a second stage, a classifier is applied to the residuals with the aim of determining the leak location. The classifier is trained with data generated by simulation of the WDN under different leak scenarios and uncertainty conditions. Several classification approaches are considered and compared. The proposed methodology is tested both using synthetic and experimental data with real WDNs of different sizes. The comparison with the current approaches shows a performance improvement.

1 Introduction

The traditional approach to leakage control is a passive one, whereby the leak is repaired only when it becomes visible. Recently developed acoustic instruments [17] allow to locate also invisible leaks, but unfortunately, their application over a large-scale water network is very expensive and time-consuming. A viable solution is to divide the network into District Metered Areas (DMA), where the *flow* and the *pressure* at the input are measured [19, 25], and to maintain a permanent leakage monitoring: leakages in fact increase the flow and decrease the pressure head at the DMA entrance. Various empirical studies [18, 31] propose mathematical models to

Adrià Soldevila, Sebastian Tornil-Sin, Joaquim Blesa, Rosa M. Fernandez-Canti, and Vicenç Puig
Research Center for Supervision, Safety and Automatic Control (CS2AC), Rambla Sant Nebridi, s/n, 08022 Terrassa (Spain) e-mail: {adria.soldevila,rosa.mari.fernandez}@upc.edu

Sebastian Tornil-Sin, Joaquim Blesa and Vicenç Puig are also at
Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Carrer Llorens Artigas, 4-6, 08028 Barcelona (Spain) e-mail: {stornil,jblesa,vpuig}@iri.upc.edu

describe the leakage flow with respect to the pressure at the leakage location. Best practice in the analysis of DMA flows consists in estimating the leakage when the demand is minimum. This typically occurs at night, when the customers demand is low and the leakage component represents a great percentage of the pipe flow [25]. Therefore, practitioners monitor the DMA or groups of DMAs for detecting, locating and estimating the leakage level by analyzing the minimum night flow [25]. However, leakage detection may not be easy, because of unpredictable variations in consumer demands and measurement noise, as well as long-term consumption trends and seasonal effects.

Several works have been published dealing with leak location methods for WDN (see [25] and references therein). For example, in [8], a review of transient-based leak detection methods is offered. In [35], a method is proposed to identify leaks by using blind spots based on the analysis of acoustic and vibrations signals [15] toghether with models of buried pipelines which allow the prediction of wave velocities [21]. In [34], Genetic Algorithms were proposed to solve an optimization problem which quantify and locate water losses. More recently, [20] have developed a method to locate leaks and estimate its outflow by using Support Vector Machines (SVM) that analyzes data obtained by a set of pressure sensors of a pipeline network. Another set of methods is based on inverse transient analysis [9, 16]. The main idea is to analyze the pressure data collected over the occurrence of transitory events by means of the minimization of the difference between the observed and the calculated parameters. In [13, 14], it is shown that unsteady-state tests can be used for pipe diagnosis and leak detection. The transient-test based methodologies use the equations for transient flow in pressurized pipes in frequency domain and then, information about pressure waves is taken into account too. More recently, the use of k-Nearest Neighbors (k-NN) and neuro-fuzzy classifiers for leak localization purposes has been proposed by [29] and [33].

Model-based leak detection and isolation techniques have also been studied starting with the seminal paper of [24] which formulates the leak detection and localization problem as a least-squares parameter estimation problem. Unfortunately, the parameter estimation of water network models is not an easy task [28]. The problem of leak localization in WDNs can be addressed as a particular case of Fault Detection and Isolation (FDI) in dynamic systems [2]. DMA hydraulic bevahior is described by a non-linear model expressed as set of algebraic equations with no explicit solution that can only be solved using numerical methods as the one proposed by [32]. This limits the applicability of most model-based FDI approaches that require to transform or manipulate the model to generate a set of residuals with the desired FDI specifications. Thus, only primary (direct) residuals could be generated that are sensitive to more than one leak because DMA typically present a dense mesh of highly interconnected pipes. This fact additionally to the reduced number of sensors make the isolation task difficult. For this reason specific model fault diagnosis methods for leak localization should be developed. A first contribution in this line can be found in [22] and [23] where a model-based method that relies on pressure measurements and leak sensitivity analysis is proposed. This method consists in computing on-line residuals, i.e. differences between the measurements and

their estimations obtained using the hydraulic network model, and defining respective thresholds that take into account the modeling uncertainty and the noise. When some of the residuals violate their threshold, the leak signature is matched with a leak sensitivity matrix to determine which of the possible leaks is present. Although this approach is efficient under ideal conditions, its performance decreases due to the nodal demand uncertainty and measurement noise. This method has been improved by [5] taking into account an analysis along a time horizon. This work presents a comparison of several leak isolation methods. It must be noticed that in cases where flow measurements are available, leaks could be detected easily since it is possible to establish simple mass balance in the pipes. See for example the work of [26] where a methodology to isolate leaks is proposed by using fuzzy analysis of the residuals. This method calculates the residuals between the flow measurements and their estimation using a model without leaks. Although the use of flow measurements is feasible in large water networks, this does not occurs when there is a dense mesh of pipes with only flow measurements at the entrance of each DMA. In this situation, water companies consider as a feasible approach the possibility of installing some pressure sensors inside the DMAs, because they are cheap and easy to install and maintain.

In this chapter, a new model-based approach for leak localization in WDNs using pressure models and classifiers is presented. This methodology is intended to be used after the leak has been detected by means of the analysis of the night DMA water demands [25], and after the application of the validation and reconstruction methodology described by [10] to the sensors used for leak localization. Following a model-based methodology successfully tested in [22] and [23], a pressure model of the considered WDN is used in a first stage to compute residuals that are indicative of leaks. In a second stage, a classifier is applied to the obtained residuals with the aim to determine the leak location. This on-line scheme relies on a previous off-line work in which the network model is obtained and the classifier is trained with data generated by extensive simulations of the network. These simulations consider three types of uncertainties: leaks with different magnitudes in all the nodes of the network, differences between the estimated and real consumer water demands and noise in pressure sensors. The underlying idea is to obtain a classifier able to distinguish the leak location independently of the unknown real leak magnitude and the presence of uncertainties associated to the water demands and the pressure measurements.

2 Background and Motivation

2.1 Principle of Model-Based Leak Location Approaches

Model-based approaches aim to locate leaks in a water distribution network by comparing pressure measurements with their estimations obtained by using the hydraulic

network model. Usually, this methodology is used for locating leaks within a given leak size range defined by the water network management company. The minimum size is related to the sensor resolution and modelling/demand uncertainty, and the maximum size is defined as the value such that the leak behaves as a burst such that it can be seen in the street. Model-based leak localization methods are based on comparing the monitored pressure disturbances caused by leaks at certain inner nodes of the DMA network with the theoretical pressure disturbances caused by all potential leaks obtained by using its respective model [23]. This comparison uses the residual vector $r \in \mathbb{R}^{n_s}$, obtained from the difference between the measured pressure at DMA inner nodes $p \in \mathbb{R}^{n_s}$ and the pressure at these nodes calculated by using the network model considering a leak-free scenario $\hat{p}_o \in \mathbb{R}^{n_s}$, i.e.

$$r(t) = p(t) - \hat{p}_o(t). \tag{1}$$

The dimension of the residual vector r, n_s, depends on the number of inner pressure sensors installed in the DMA. In recent years, some optimal sensor placement algorithms have been developed to determine where the pressure head sensors should be installed inside the DMA with minimum economical costs (number of sensors), a suitable performance regarding leak localization is guaranteed, see [22], [7], [27] among others.

The number of potential leaks, $f \in \mathbb{R}^{n_n}$, is considered to be equal to the number of DMA nodes n_n, since from the modeling point of view, as proposed by [22] and [23], leaks are assumed to occur in these locations.

2.2 Limitations of Sensitivity Analysis Approaches

Most model-based leak localization approaches rely on the sensitivity-to-leak analysis [22, 23] where the theoretical pressure disturbances caused by all potential leaks are stored in the leak sensitivity matrix $\Omega \in \mathbb{R}^{n_s \times n_n}$ (with as many rows as DMA inner pressure sensors, n_s, and as many columns as potential leaks in all nodes n_n). Then, leak isolation is based on matching the residual vector (1) with the columns of the sensitivity matrix by using some metrics as for example, the correlation or the angle (see [5] for details). The leak sensitivity matrix can be mathematically formalized as follows

$$\Omega = \begin{pmatrix} \frac{\partial r_1}{\partial f_1} & \cdots & \frac{\partial r_1}{\partial f_{n_n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_{n_s}}{\partial f_1} & \cdots & \frac{\partial r_{n_s}}{\partial f_{n_n}} \end{pmatrix}, \tag{2}$$

where each element $\Omega_{i,j}$ measures the effect of the leak f_j in the residual r_i associated to the pressure at node i. In practice, it is extremely difficult to calculate Ω analytically because a water network is a large scale multivariable non-linear system which equations can only be solved numerically as discussed in the intro-

duction. Thereby, the sensitivity matrix is generated by simulation of the network model and evaluating the sensitivity $\Omega_{i,j}$ as

$$\Omega_{i,j} = \frac{\hat{p}_{i,f_j} - \hat{p}_{i,0}}{f_j}, \tag{3}$$

where $\hat{p}_{i,f_j}$ is the predicted pressure in the node when a nominal leak f_j is forced in node j and $\hat{p}_{i,0}$ is the predicted pressure associated with the sensor i under a scenario free of leaks [22]. Then, the sensitivity matrix is obtained by repeating this process for all n_n potential faults .

An important drawback of the leak sensitivity approach is that the practical evaluation of (3) depends on the nominal leak f_j [3, 4]. If the real leak size is different from the nominal one, the real sensitivity will be different from the one computed using (3). Moreover, the sensitivity is also affected by the nodal demand uncertainty [11] since this demand is not measured but estimated using historical records of water consumption and using the aggregated DMA consumption pattern. These uncertainties will deteriorate the leak localization results obtained by using the sensitivity approach. The approach proposed in this chapter aims to overcome these difficulties.

3 Proposed Method

3.1 Basic Architecture and Operation

The method for on-line leak localization proposed in this chapter relies on the scheme depicted in Figure 1, and it is based on computing pressure residuals and analyzing them with a classifier. The hydraulic model is built using the Epanet hydraulic simulator[1] by considering the DMA structure (pipes, nodes and valves) and network parameters (pipe coefficients). After the corresponding calibration process using real data, it is assumed that the hydraulic model is able to represent precisely the WDN behavior. However, it must be noted that the model is fed with estimated water demands in the nodes $(\hat{d}_1, \cdots, \hat{d}_{n_n})$. In practice, nodal demands $(d_1, \cdots, d_{n_n})$ are not measured (except for some particular consumers where Automatic Metering Readers (AMRs) are available) and are typically obtained by the total measured DMA demand $\tilde{d}_{WDN}$ and distributed at nodal level using historical consumption records. Hence, the residuals are not only sensitive to leaks but also to differences between the real demands and their estimated values. Additionally, pressure measurements are subject to the effect of sensor noise v and this also affects the residuals. Taking all these effects into account, the classifier must be able to locate the real leak present in the WDN, that can be in any node and with any (unknown) magnitude, while being robust to the demand uncertainty and the measurement noise. Finally, the operation of the network is constrained by some boundary conditions c

[1] https://www.epa.gov/water-research/epanet

(such as the position of internal valves, reservoir pressures and flows) that are known (measured) and must be taken into account in the simulation and can also be used as inputs for the classifier.

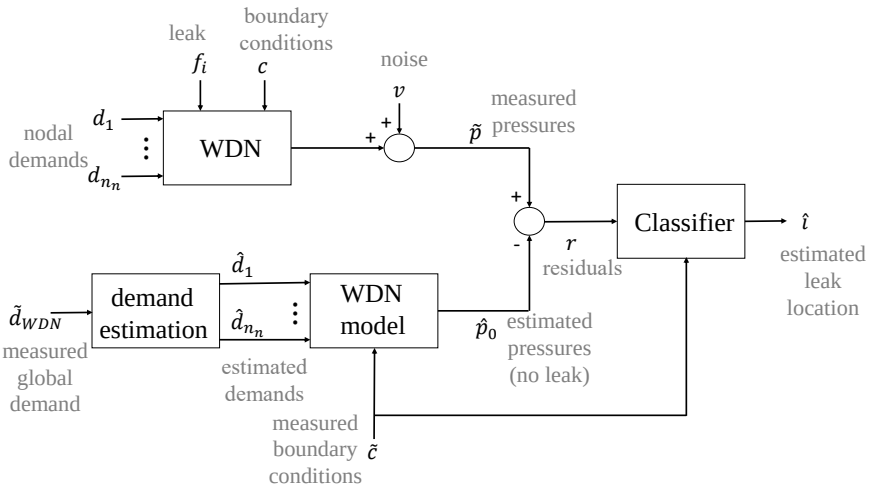

Fig. 1 Leak localization scheme

3.2 Methodology Overview

The exploitation of the architecture presented Figure 1 relies on a methodology that distinguishes several off-line and on-line procedures.

3.2.1 Off-line

The application of the architecture presented in Figure 1 relies on an off-line work whose main goal is to obtain a classifier able to distinguish the potential leaks under the described uncertainty conditions. In particular, the method proposed in this chapter considers an off-line design based on the following stages:

- Modelling - A model for the WDN is obtained, calibrated and implemented with Epanet. The model is basically built by taking into account the network structure and by applying flow balance conservation and pressure loss equations (see [22] and [23] for details).

- Data generation - The model implemented is extensively used to generate data in the residual space for each possible leak and for different operating and uncertainty conditions.
- Classifier training and evaluation - The classifier is first trained with a subset of the initial data set, then it is applied to testing data in order to estimate its performance.

The data generation stage is critical since the availability of representative data is a necessary condition for obtaining a good classifier. Since the amount of data collected from the real monitored WDN is limited, a way to obtain a complete training data set is by using the hydraulic simulator. Hence, training and testing data are generated by applying the scheme depicted in Figure 2, which is similar to the one presented in Figure 1 but with the main difference of substituting the real WDN by a model that allows to simulate the WDN not only in absence but also in presence of faults.

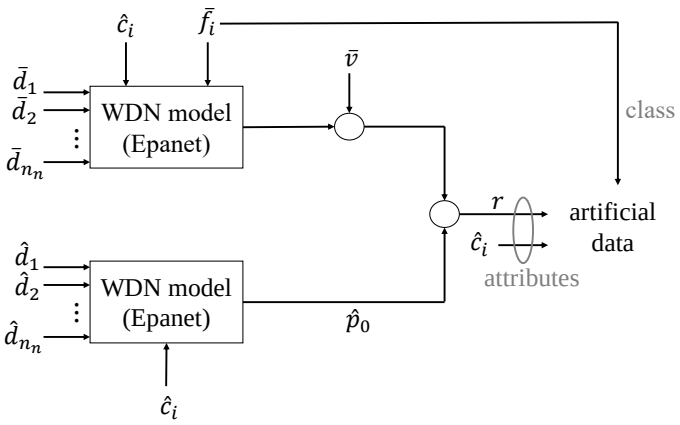

Fig. 2 Data generation scheme

The presented scheme is exploited in order to:

- Generate data for all possible leak locations, i.e. for all the different nodes in the WDN ($\bar{f}_i, \quad i = \{1,2,...,n_n\}$).
- Generate data for each possible leak location with different leak magnitudes within a given range ($\bar{f}_i \in [f_i^-, f_i^+]$).
- Generate sequences of demands and boundary conditions $\hat{c}_i$ that correspond to realistic typical daily evolution in each node.

- Simulate differences between the real demands and the estimations computed by the demand estimation module ($(\bar{d}_1,...,\bar{d}_{n_n}) \neq (\hat{d}_1,...,\hat{d}_{n_n})$).
- Take into account the measurement noise in pressure sensors, by generating synthetic Gaussian noise ($\bar{v}$).

The artificial data obtained from simulations is divided into training and testing sets. The training stage is based on a learning procedure where the input is the (labeled) training data set and the result is a classifier that must be able to correctly classify new data instances into the correct class. The generalization ability of the obtained classifier is checked in the validation stage, in which the performance indexes are computed for the testing data set.

The details of the training stage are particular of the type of classifier used. The results presented have been obtained by using two different well-known classifiers: the k-Nearest Neighbor (k-NN) classifier [1], which is non-parametric, and the Bayesian classifier, which is parametric. The details about the training of both classifiers will be provided in the next subsections.

The evaluation of classifiers normally relies on the use of the *confusion matrix* Γ, that summarizes the results obtained when the classifier is applied to the testing data set. The confusion matrix is a square matrix with as many rows and columns as nodes of the network (potential leak locations), when it is applied to the leak localization problem with the associated terminology. Each coefficient $\Gamma_{i,j}$ indicates how many times a leak in node i is recognized as a leak in node j. Table 1 illustrates the concept of the confusion matrix applied to leak localization.

Table 1 Confusion matrix Γ

	$\hat{f}_1$	$\cdots$	$\hat{f}_i$	$\cdots$	$\hat{f}_{n_n}$
f_1	$\Gamma_{1,1}$	$\cdots$	$\Gamma_{1,i}$	$\cdots$	Γ_{1,n_n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
f_i	$\Gamma_{i,1}$	$\cdots$	$\Gamma_{i,i}$	$\cdots$	Γ_{i,n_n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
f_{n_n}	$\Gamma_{n_n,1}$	$\cdots$	$\Gamma_{n_n,i}$	$\cdots$	Γ_{n_n,n_n}

In case of a perfect classification, Γ is diagonal, with $\Gamma_{i,i} = m$, for all $i = 1,\cdots,n_n$, being m the size of the testing data set. In practice, non-zero coefficients will appear outside the main diagonal. For a leak in node i, the coefficient $\Gamma_{i,i}$ indicates the number of times that the leak is correctly identified as $\hat{f}_i$, while $\sum_{j=1}^{n_n} \Gamma_{i,j} - \Gamma_{i,i}$ indicates the number of times that is wrongly classified. The overall accuracy (Ac) of the classifier is defined as

$$Ac = \frac{\sum_{i=1}^{n_n} \Gamma_{i,i}}{\sum_{i=1}^{n_n} \sum_{j=1}^{n_n} \Gamma_{i,j}}. \tag{4}$$

3.2.2 On-line

Once the classifier has been trained and validated, it can be used on-line to localize leaks. According to Section 3.1, the classifier can be directly used to detect leaks based on the instantaneous values of the computed residuals.

However, this strategy may provide limited results if there is a high level of uncertainty. The use of a temporal reasoning that takes into account not only the instantaneous values of the residuals but all the values within a time horizon is suggested as already suggested in [6]. This idea is implemented in different forms depending on the type of classifier that is used, details are provided in the next subsections.

3.3 k-NN Classifier Implementation

3.3.1 The *k*-NN Classifier

One of the well accepted and established methods for classification is the *k*-NN algorithm [1], which is available in most numerical packages (e.g. Matlab, R, etc.). Its basic version works as follows. When a new data realization has to be classified, the distance (typically, the Euclidean distance is used, but many other options are available) to all the instances of the training data set is computed. Then, the k nearest neighbors are selected and a voting procedure is applied, where each neighbor votes for its own class and the class with more votes is chosen. The process is illustrated in Figure 3, where a value $k = 3$ is used and the new data instance is associated to the class C_3 since two of the three minimal distances correspond to training instances of that class. The value for k is typically bigger than one to improve the robustness against outliers and it must be smaller than the minimum number of instances of a single class from the training data set. The *k*-NN classifier is said to be a lazy classifier since the training procedure is limited to the storage of the training set and all the computations are deferred until the classification process is performed.

3.3.2 Time Reasoning

If the uncertainty in the demands, the leak magnitude or the noise level are large then the direct application of the classifier can provide poor leak localization results. This also happens when other ways of evaluating the pressure residuals are used (as the ones described in Section 2). To smooth the effect of demand uncertainty, leak magnitude and noise, typically the analysis of the residuals evolution is performed in a time horizon, i.e. the values for the residuals in the last N time instants are considered [6].

A simple temporal reasoning can be based on taking into account the leak localization results provided by the classifier inside the time horizon and applying a

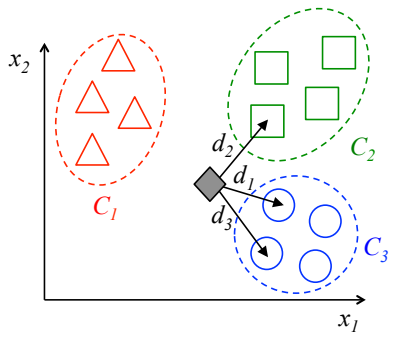

Fig. 3 The k-NN algorithm

voting scheme, concluding that the candidate leak is located in the node that more times has been selected by the classifier.

A second and more sophisticated option could be to use the information contained in the confusion matrix. Hence, at each time instant t, when the classifier is providing a leak in node j as an explanation for the values of the residuals at the current time instant t, the whole column j of the confusion matrix is stored. This column provides an estimation of the probabilities $p(f_i|\hat{f}_j)$, i.e. the probabilities of a leak at node i when the classifier indicates that the leak is at node j, according to the information available for current time instant t. Then, the sum of column vectors stored along the time horizon N is computed. In the obtained vector, the position of the coefficient with highest value indicates the most probable leak according to the information provided by the data in the time horizon $[t-N+1,t]$.

3.4 Bayesian Classifier Implementation

3.4.1 Bayesian Classification

Assume that we have a finite set of possible leak situations (i.e. classes), f_i, $i = 1,...,n_n$ and a finite set of measuring devices x_j, $j = 1,...,n_s$. Assume also that we have a model of the system behavior which allows the computation, at each time

instant t, of n_s residual signals r_j, $j = 1, ..., n_s$. When a leak occurs, all the residuals are activated up to some extent different from zero.

Given the residuals, the objective is to apply a Bayesian leak discrimination procedure in order to identify which leak or leaks may occur based on the observed behavior. Such a diagnosis procedure based on Bayesian reasoning is explained below.

At every time sample t, the probability of a leak occurrence is estimated as a result of the application of the Bayes Rule

$$P(f_i \mid \mathbf{r}(t)) = \frac{P(\mathbf{r}(t) \mid f_i)P(f_i)}{P(\mathbf{r}(t))}, \, i = 1, ..., n_n, \tag{5}$$

where $P(f_i \mid \mathbf{r}(t))$ is the posterior probability that the leak f_i had caused the observed residual vector $\mathbf{r}(t) = (r_1(t) \cdots r_j(t))^T$, $P(\mathbf{r}(t) \mid f_i)$ is the likelihood of the residual $\mathbf{r}(t)$ assuming that the active leak is f_i, $P(f_i)$ is the prior probability for the leak f_i, and $P(\mathbf{r}(t))$ is a normalizing factor given by the Total Probability Law,

$$P(\mathbf{r}(t)) = \sum_{i=1}^{n_n} P(\mathbf{r}(t) \mid f_i)P(f_i). \tag{6}$$

Regarding the prior probabilities, unless we have any additional information, an unprejudiced starting point is to consider all them equally probable, that is, $P(f_i) = \frac{1}{n_n}$, $i = 1, ..., n_n$. To estimate the likelihood value $P(\mathbf{r}(t) \mid f_i)$, we need to perform a previous calibration task in order to obtain the joint probability density function for each leak in the residual space, $P(\mathbf{r} \mid f_i), i = 1, ..., n_n$. The calibration stage is detailed in a next section. Note that, in contrast to standard naïve Bayesian classifiers, we do not need to assume independence between the residuals.

The application of (5) produces a set of values $P(f_i \mid \mathbf{r}(t))$, $\sum_{i=1}^{n_n} P(f_i \mid \mathbf{r}(t)) = 1$, that can be used to decide which leak is acting over the system. Note that, at each time sample t, we have information about the probability associated with each leak situation. Thus, there can be many competing leaks, each one with a different probability value. The leak with the highest posterior probability can be selected as the most likely leak, or all the leaks with a posterior probability above a pre-specified threshold can be selected as leak candidates.

3.4.2 Recursivity

The results can be improved if (5) is recursively applied, that is, if the posterior probability $P(f_i \mid \mathbf{r}(t))$ is used as the prior probability for the next time sample. This way, as long as new measurement data are available, the probabilities are updated and many of the competing leaks can be discarded.

The only drawback is that if any of the leaks takes the posterior probability value of 1 at any t, then all the remaining leaks take the 0 probability value, therefore preventing them to have a future value different from zero due to the recursive application of (5). This drawback can be easily overcomed by forcing all probabilities

to have a maximum value of, say, 0.99. When a leak f_i presents the probability $P(f_i \mid \mathbf{r}(t)) > 0.99$, we force it to be $P(f_i \mid \mathbf{r}(t)) = 0.99$ and we can force the remaining leaks to be $P(f_n \mid \mathbf{r}(t)) = \frac{1-0.99}{n_n-1}$, $n = 1,...,n_n, n \neq i$.

3.4.3 Bayesian Time Reasoning

Additionally, the results can be improved if a time horizon N is introduced. In this case, the posterior probability can be computed on the basis of the N previous time samples, that is, to compute $P(f_i \mid \mathbf{r}(t))$, we recursively can apply the following equation

$$P(f_i \mid \mathbf{r}(t-N+n)) = \frac{P(\mathbf{r}(t-N+n) \mid f_i)P(f_i \mid \mathbf{r}(t-N+n-1))}{P(\mathbf{r}(t-N+n))}, \quad i = 1,...,n_n,\ n = 1,...,N, \tag{7}$$

where an unprejudiced starting point may be $P(f_i \mid \mathbf{r}(t-N)) = \frac{1}{n_n}$, $i = 1,...,n_n$.

3.4.4 Calibration of the Probability Density Functions

Unlike the k-NN classifier, the Bayes classifier requires a more elaborated training where a joint Probability Density Function (PDF) for each leak class in the residual space, $P(\mathbf{r} \mid f_i)$, $i = 1,...,n_n$, has to be estimated.

The first step is to decide the probability family. The Law of Large Numbers states that most situations lead to a Gaussian probability density function if the number of samples is high enough. Several tests can be applied to the residual values to assess if they are Gaussian distributed or not. For instance, we can apply the well-known Kolmogorov-Smirnov [12] or the Anderson-Darling [30] tests, among others.

Figure 4 shows the two leak distributions calibrated by means of Gaussian probability function. Leak 1 is better adjusted because it takes into account the cross-correlation between residuals r_1 and r_2. On the other hand, leak 2 is adjusted by assuming statistic independence between residuals r_1 and r_2 and therefore the fitting is not so accurate. Note also that other probability distribution families different from Gaussian could be used, including multimodal and non-parametric distributions.

4 Case Studies

In this section, two DMA case studies of increasing size and complexity (Hanoi and Nova Icària) are introduced to assess the performance of the proposed methodology.

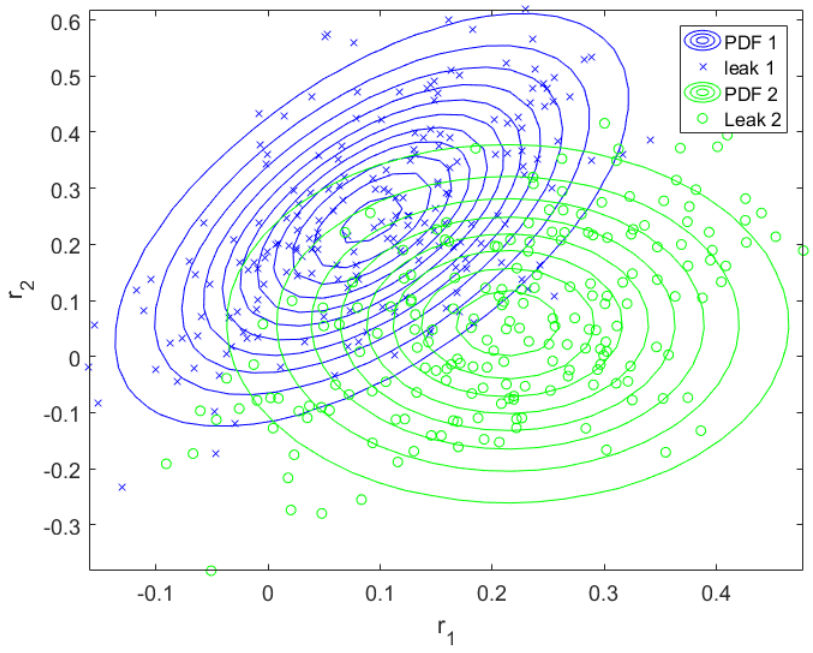

Fig. 4 Calibration for leaks 1 and 2

For these DMAs, leaks are considered in any of the demand nodes. The known variables are the input pressures and flows of the networks (reservoir boundary conditions) and some pressures at the inner nodes of the DMAs where sensors would be located (see [7, 27] and [3] for details about optimal sensor location). It is considered that the demand pattern is known for all demand nodes but with some uncertainty as proposed by [11]. The leak magnitude is assumed to be unknown but bounded by a known interval (minimum and maximum leak magnitudes). Finally, noise in pressure sensors is considered too.

For the two DMAs, leak localization results under different uncertainty scenarios are presented and discussed. Moreover, for the second (and biggest) DMA the results of localizing a real leak are also presented.

4.1 Hanoi Case Study

The proposed methodology has been first applied to the simplified model of the Hanoi (Vietnam) DMA network, depicted in Figure 5. This model consists of 1 reservoir, 34 pipes and 31 nodes. Measurements of two inner pressure sensors placed in nodes 14 and 30 are available as considered in other works [29].

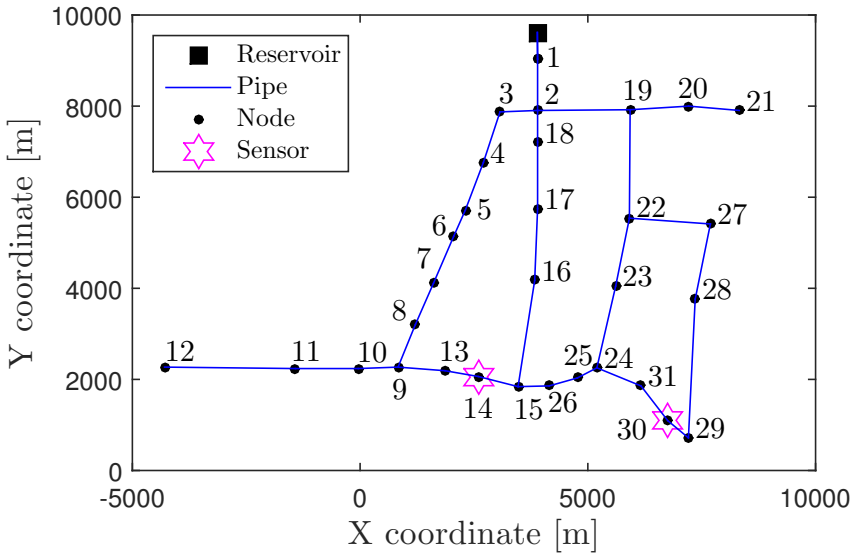

Fig. 5 Hanoi topological network

In order to illustrate the performance of the proposed methodology, four different studies have been carried out under the following particular conditions:

- A leak size uncertainty study considering a leak range between 25 and 75 l/s (0.84 and 2.51 % of the total amount of water demanded, which is 2991 l/s).
- A study considering noise in pressure measurements with an amplitude of ± 5 % of the mean value for all pressure residuals.
- A demand uncertainty study considering an uncertainty of ± 10 % of the nominal demand node values.
- A study considering that all the three uncertainties previously defined are simultaneously present in the DMA.

For each study, two complete data sets have been generated for each node (potential leak locations), one for training purposes and the other one for testing the leak localization performances. Each set used for testing, associated to a leak at a given node, is called a leak scenario. The variables conforming the data are the input flow $\tilde{d}_{DMA}$ and the two residuals r_1 and r_2 associated to pressure measurements in nodes 14 and 30, respectively. The feature space used as input for the classifier is represented in Figure 6. The sampling time used in the simulations is 10 minutes, but hourly average values of variables are used to improve the leak location performance. Different daily input flow patterns have been simulated as the one depicted in Figure 7. Accordingly to the scheme presented in Figure 1, the pressure residuals

have been obtained by means of a WDN simulator (Epanet model of the network) where the uncertainties described above have been considered.

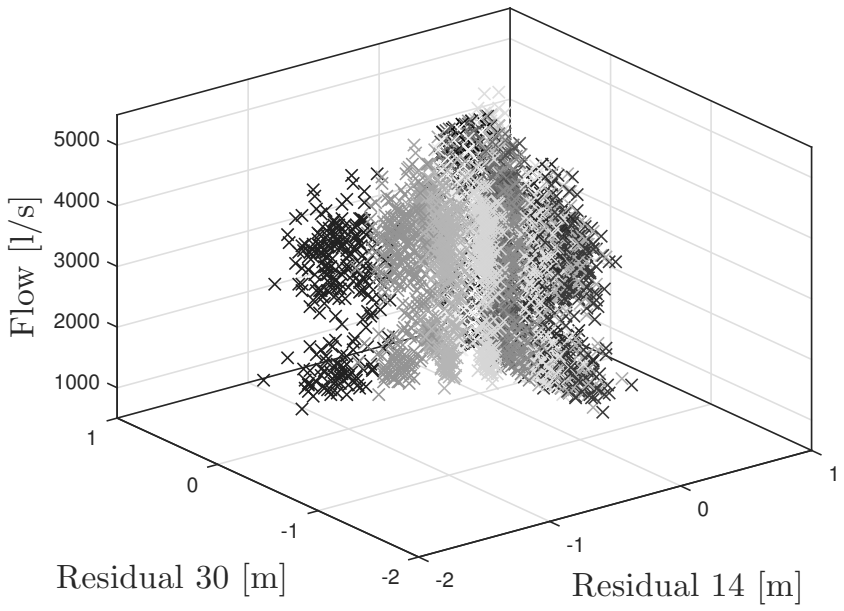

Fig. 6 Hanoi residual space with uncertainties (each color represents a leak in a different location)

In order to determine whether the three classifier inputs (r_1, r_2 and $\tilde{d}_{DMA}$) follow a Gaussian distribution, a one-dimension Kolmogorov-Smirnov test on a training data set of 480 samples (for each of the 31 leak nodes) has been performed. As a result, the three inputs can be considered Gaussian distributed for a significance level of 3%.

The results obtained by the proposed method in the four different studies have been compared to the ones obtained by using the leak-sensitivity analysis with the angle metrics proposed by [5] and summarized in Section 2. For the Angle method only the two residuals are used because the flow measurement has a great value and tends to reduce the effect of residuals in the diagnosis, thus resulting in worse results. The sensitivity matrix (2) has been computed using (3) and by considering nominal leak conditions in every demand node ($\bar{v} = 0$, $\bar{d} = \hat{d}$ and $f_i = 50$ [l/s] $i = 1, ..., n_n$). The results obtained by using the Angle method and the two proposed methods, in both cases considering only one sample ($N = 1$) and the equivalent number of samples of one day ($N = 24$) in the leak location diagnosis are summarized in Table 2. The values presented in the table correspond to the overall accuracy Ac defined in (4).

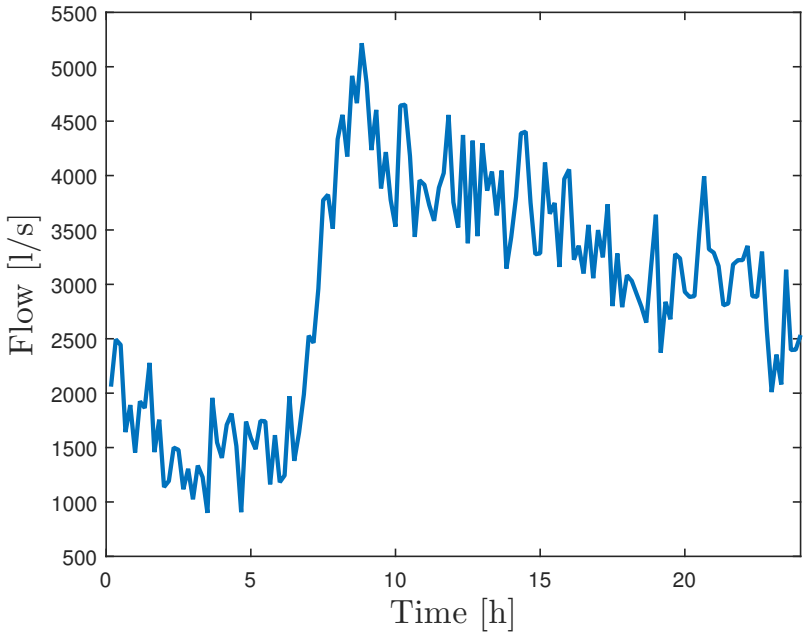

Fig. 7 Example of a daily Hanoi flow pattern

Table 2 Accuracy results in Hanoi network

Study	$N = 1$			$N = 24$		
	k-NN	Bayes	Angle	k-NN	Bayes	Angle
Leak uncertainty	60.21	83.60	76.61	77.41	83.87	77.41
Noise in measurements	69.62	83.19	73.79	83.87	83.87	70.96
Demand uncertainty	31.18	39.11	41.39	58.06	45.16	64.51
All together	32.12	48.25	36.96	74.19	83.87	54.83

As it can be seen, the three methods provide good performance in the leak uncertainty case because of the linear directional variation of most of the residuals for this kind of uncertainty [3]. It must be noted that in the case when only demand uncertainty is considered, the classifier-based methods perform worse than when all the uncertainties are considered together, this happens because the leak uncertainty spread the residual data providing a better separation (and for the Bayesian classifier the distribution tends to be more Gaussian).

When the time horizon and recursivity described in Section 3.4.1 are applied, it can be seen that there is an improvement in the performance achieved in all uncertainty cases (except for the case of the noise uncertainty for the Angle method, where the full performance is achieved since the first sample, and then fluctuates within the time horizon around the same values).

The effect of the horizon length N in the performance (Accuracy) for the three studied methods is also analyzed using the last study (to create the figures an extended data set, ten times larger, has been used). The results for the k-NN classifier are shown in Figure 8, the results for the Bayesian classifier are shown in Figure 9, and the results for the Angle method are shown in Figure 10. The term"node relaxation" refers to the number of nodes in topological distance between the node with the real leak and the node where the classifier predict the leak for which the diagnosis is still considered correct.

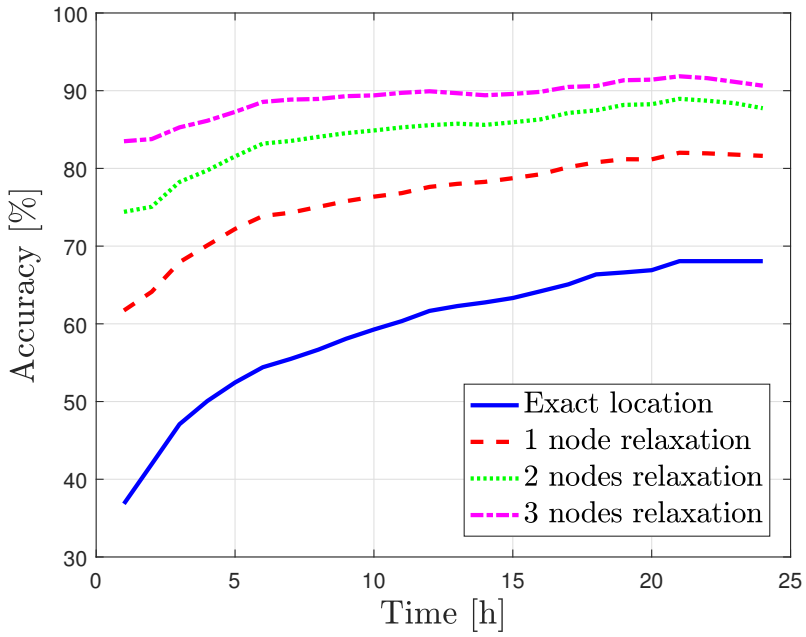

Fig. 8 Accuracy results over a time horizon for the k-NN classifier in Hanoi network

As expected, the accuracy increases with the time horizon length N. It can be observed that it reaches a steady state value when N is around twenty hours. This result justifies the use of a time horizon corresponding to one day and it agrees with the results already presented by [5].

Finally, Figure 11 shows a comparison of the three studied methods by using a different performance indicator, the Average Topological Distance, which is the minimum distance in nodes between the node candidate and the node where the leak exists.

The results show the good performance of the classifiers, especially the Bayesian classifier, which works better than the k-NN classifier when the data has a clear distribution (if not, the k-NN performs better as it can be seen in the Table 2 for the

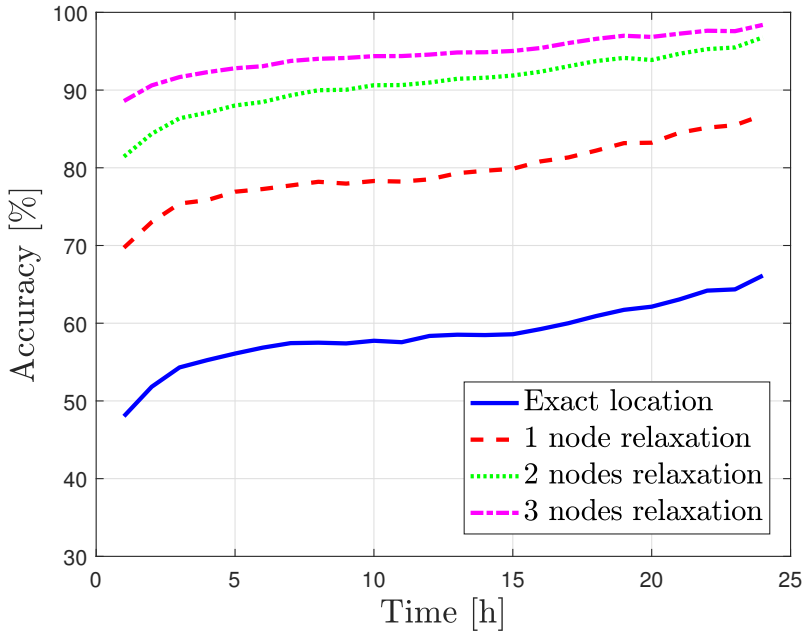

Fig. 9 Accuracy results over a time horizon for the Bayesian classifier in Hanoi network

demand uncertainty case), and also has a better reasoning over time. Also, in Figure 11, it can be seen that the Bayesian classifier tends to point a closer class when it fails than the k-NN classifier, but it can increase its performance at that point by choosing a bigger k value. To sum up, the Bayesian classifier should be used when the classes present a Gaussian distribution, and the k-NN classifier otherwise.

4.2 Nova Icària Case Study

The Nova Icària network, shown in Figure 12, is one of the DMA networks of the Barcelona WDN. This network consists in 1520 nodes, 1646 pipes, 2 reservoirs and 2 valves, each one after the reservoirs with the aim of maintaining a certain pressure level. Inside the network, the pressures measured by five sensors installed in nodes 3, 4, 5, 6 and 7 are known, together with the flow entering the DMA and the set points for the valves.

As in the previous network examples, some leak localization studies have been carried out by simulation. But additionally, a real case is studied. For this real case, experimental data captured under normal network operation and under the presence of a real leak is used. The leak was created by the water company that operates

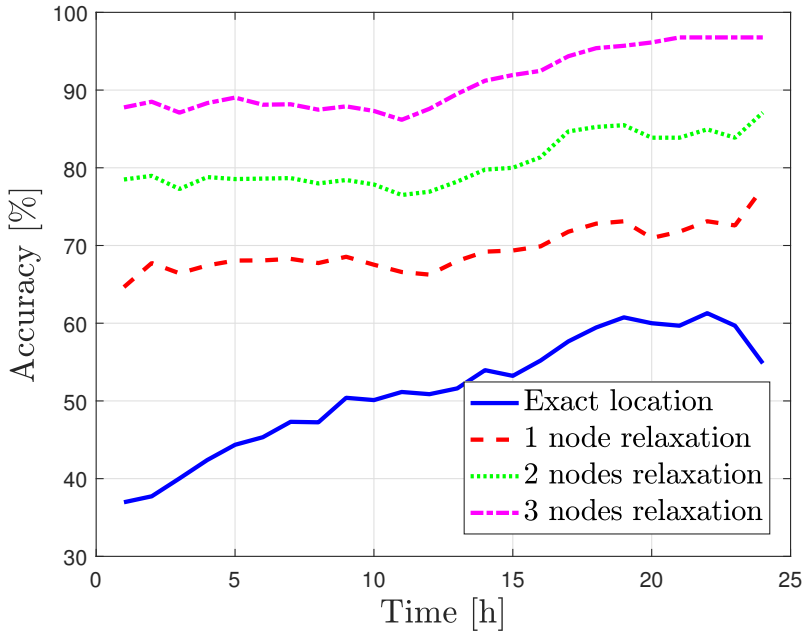

Fig. 10 Accuracy results over a time horizon for the Angle method in Hanoi network

the network by opening a fire hydrant. The experiment took place on December 20, 2012 at 00:30 h and lasted around 30 hours with a leak size about 5.6 [l/s], being the total demand of water in the range between 23.5 and 78 [l/s] approximately. Additionally, data captured in a normal operation scenario of five days before the leak scenario was also obtained. For more details see [23]. The sampling time of all data sensors is 10 minutes. In order to decrease the effect of uncertainties, the average value of every six samples has been considered every hour, i.e. 30 and 120 hourly samples are available for the leak and normal operation scenarios. An accurate Epanet model of the network and node demand estimations were provided as well.

First, the system has been simulated considering the operating conditions of the fault-free scenario (input flow, boundary conditions and demand distributions). The differences between the 120 hourly samples of the five inner pressure sensors and the pressures estimated by the hydraulic model have been used to estimate the real uncertainty of the network (demand uncertainty, modeling errors and noise in the measurements). On the other hand, nominal hourly leak residuals $r_i^0(t)$, $i = 1,\ldots,n_n$, $t = 1,\ldots,24$ have been computed as the difference of the estimated pressures in the five inner sensors in a leak scenario and the ones estimated in the normal operation.

A k-NN classifier (with $k = 3$) has been trained for leak localization and validated. The inputs of the classifier are: the five pressure residuals, the flow that enters

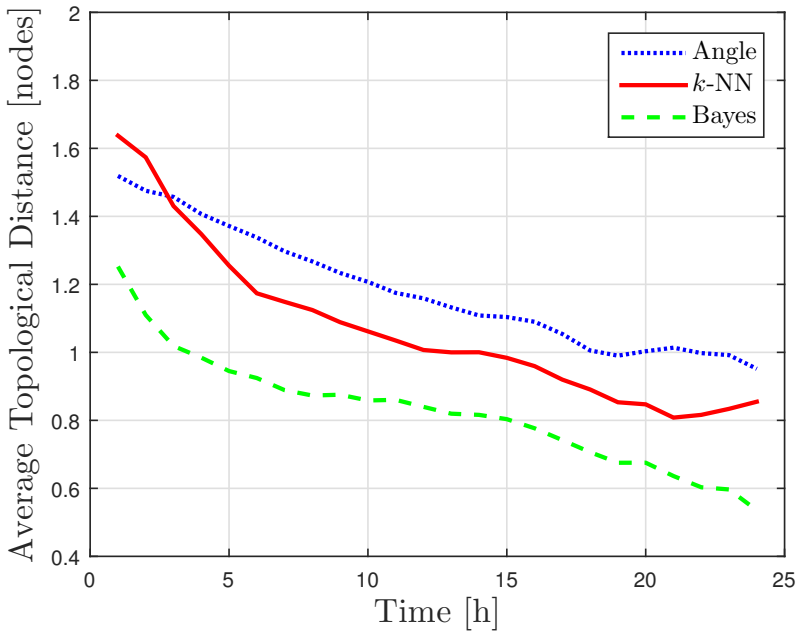

Fig. 11 Average topological distance results in a time horizon in Hanoi network

the DMA and the two set points of the valves. The data used in the training and testing stages are the 24 samples of nominal hourly residuals directly and adding the real uncertainty (120 samples): 96 samples for training and 48 for validation. The same training data sets generated are used to calibrate the PDFs for the Bayesian classifier.

Figure 13 shows the result of the two proposed methods after applying 24 hourly samples: the k-NN classifier indicates that the leak is in node 3 while the real leak is in node 996, which means that the topological distance is 13 nodes, and the geographical linear distance is around 184 meters. For the Bayesian classifier, the node candidate is 403 which has a topological distance of 10 nodes and a geographical linear distance of 183 meters. As a comparison, the application of correlation method [23] provides as node candidate the node 1036 (this result is also shown in Figure 13), which is at a distance of 17 nodes and 222 meters of the real leak location.

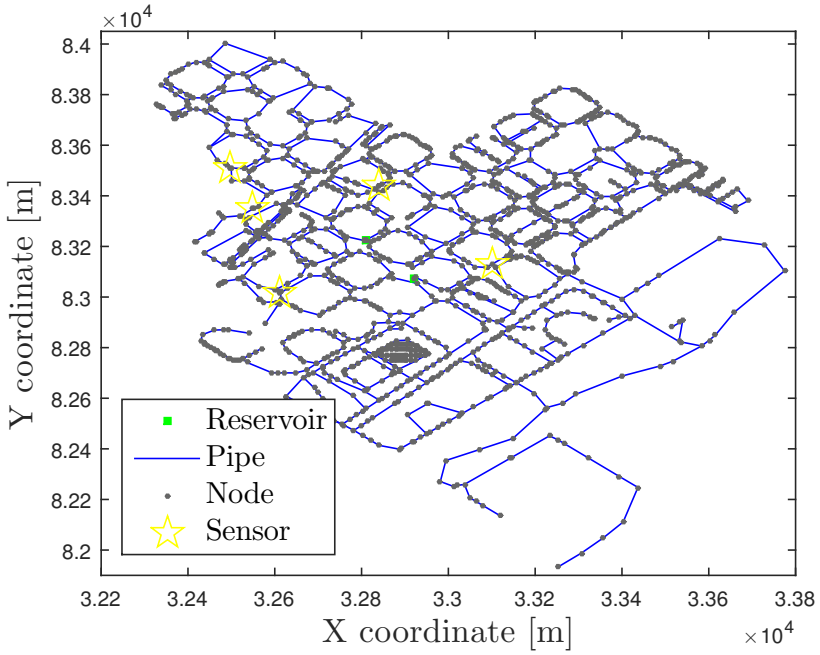

Fig. 12 Nova Icària topological network

5 Conclusion

This chapter has proposed a new method for leak localization in WDNs that combines the use of pressure models with classifiers. A model of the considered WDN is used in a first stage to compute pressure residuals that are indicative of leaks. In a second stage, a classifier is applied to the obtained residuals with the aim of determining the leak location. This on-line scheme relies on a previous off-line work in which the model is obtained and the classifier is trained with data generated in extensive simulations of the network under different leak conditions. These simulations consider leaks with different magnitudes in all the nodes of the network, differences between the estimated and consumer real water demands and noise in pressure sensors. The proposed method has been compared with a previous leak localization method described in the literature through their application to two DMA case studies of different size and complexity obtaining satisfactory results.

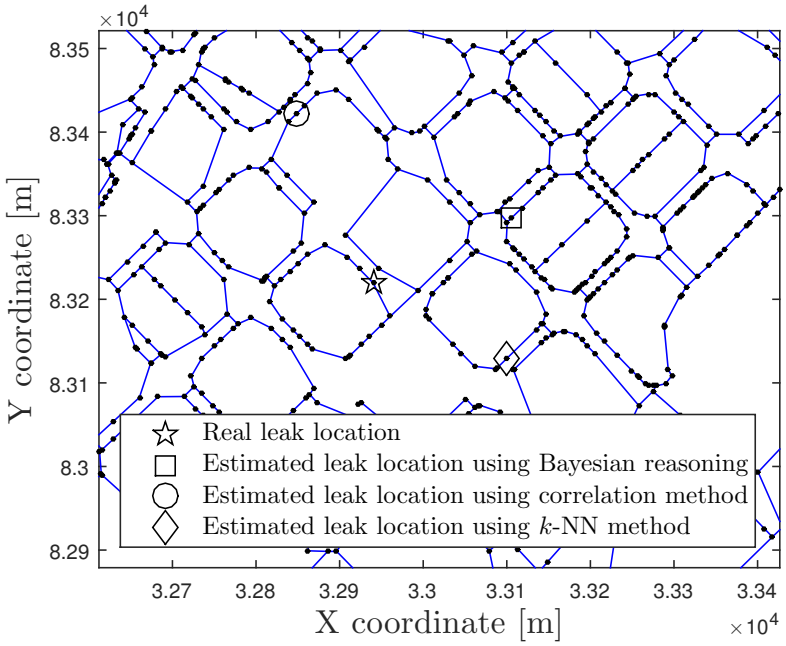

Fig. 13 Comparison of different leak location methods in Nova Icària network

6 Acknowledgment

This work has been funded by the Spanish Ministry of Economy and Competitiveness (MINECO) and FEDER through the projects ECOCIS (ref. DPI2013-48243-C2-1-R) and HARCRICS (ref. DPI2014-58104-R) and through the grant IJCI-2014-2081,by the European Commission through contract EFFINET (ref. FP7-ICT2011-8-31 8556) and by the Catalan Agency for Management of University and Research Grants (AGAUR) through the grants FI-DGR 2015 (ref. 2015 FI_B 00591) and 2014PDJ00102.

References

1. Alpaydin, E.: Introduction to Machine Learning. MIT Press (2010)
2. Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M.: Diagnosis and Fault-tolerant Control. pringer-Verlag, Berlin/Heidelberg. (2006)
3. Blesa, J., Nejjari, F., Sarrate, R.: Robust sensor placement for leak location: Analysis and design. Journal of Hydroinformatics **18**(1), 136 – 148 (2016)
4. Blesa, J., Puig, V., Saludes, J.: Robust identification and fault diagnosis based on uncertain multiple input multiple output linear parameter varying parity equations and zonotopes. Journal of Process Control **22**(10), 1890–1912 (2012)

5. Casillas, M.V., Garza-Castañón, L.E., Puig, V.: Extended-horizon analysis of pressure sensitivities for leak detection in water distribution networks. In: 8th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, pp. 570–575. Elsevier (2012)
6. Casillas, M.V., Garza-Castañón, L.E., Puig, V.: Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities. Journal of Hydroinformatics (2013 (in press))
7. Casillas, M.V., Puig, V., Garza-Castañón, L.E., Rosich, A.: Optimal sensor placement for leak location in water distribution networks using genetic algorithms. Sensors **13**(11), 14,984–15,005 (2013)
8. Colombo, A.F., Lee, P., Karney, B.W.: A selective literature review of transient-based leak detection methods. Journal of Hydro-environment Research pp. 212–227 (2009)
9. Covas, D., Ramos, H.: Hydraulic transients used for leak detection in water distribution systems. In: 4th International Conference on Water Pipeline Systems, BHR Group, pp. 227–241 (2001)
10. Cugueró-Escofet, M.À., García, D., Quevedo, J., Puig, V., Espin, S., Roquet, J.: A methodology and a software tool for sensor data validation/reconstruction: Application to the Catalonia regional water network. Control Engineering Practice **49**, 159 – 172 (2016)
11. Cugueró-Escofet, P., Blesa, J., Pérez, R., Cugueró-Escofet, M.À., Sanz, G.: Assessment of a leak localization algorithm in water networks under demand uncertainty. IFAC Proceedings Volumes (IFAC-PapersOnline) **48**(21), 226–231 (2015)
12. Daniel, W.W.: Applied nonparametric statistics. PWS-Kent Boston (1990)
13. Ferrante, M., Brunone, B.: Pipe system diagnosis and leak detection by unsteady-state tests. 1. Harmonic analysis. Advances in Water Resources **26**(1), 95–105 (2003)
14. Ferrante, M., Brunone, B.: Pipe system diagnosis and leak detection by unsteady-state tests. 2. Wavelet analysis. Advances in Water Resources **26**(1), 107–116 (2003)
15. Fuchs, H.V., Riehle, R.: Ten years of experience with leak detection by acoustic signal analysis. Applied Acoustics (33), 1–19 (1991)
16. Kepler, A., Covas, D., Reis, L.: Leak detection by inverse transient analysis in an experimental *PVC* pipe system. Journal of Hydroinformatics **13**(2), 153–166 (2011)
17. Khulief, Y., Khalifa, A., Mansour, R., Habib, M.: Acoustic detection of leaks in water pipelines using measurements inside pipe. Journal of Pipeline Systems Engineering and Practice **3**(2), 47–54 (2012)
18. Lambert, A.: What do we know about pressure:leakage relationships in distribution systems? In: IWA Conference System Approach to leakage control and water distribution system management. Brno, Czech Rebublic. (2001)
19. Lambert, M.F., Simpson, A.R., Vítkovský, J.P., Wang, X.J., Lee, P.J.: A review of leading-edge leak detection techniques for water distribution systems. In: 20th AWA Convention, Perth, Australia (2003)
20. Mashford, J., De Silva, D., Marney, D., Burn, S.: An approach to leak detection in pipe networks using analysis of monitored pressure values by support vector machine. In: Third International Conference on Network and System Security, pp. 534–539 (2009)
21. Muggleton, J.M., Brennan, M.J., Pinnington, R.J.: Wavenumber prediction of waves in buried pipes for water leak detection. Journal of Sound and Vibration (249), 939–954 (2002)
22. Pérez, R., Puig, V., Pascual, J., Quevedo, J., Landeros, E., Peralta, A.: Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks. Control Engineering Practice **19**(10), 1157 – 1167 (2011)
23. Pérez, R., Sanz, G., Puig, V., Quevedo, J., Nejjari, F., Meseguer, J., Cembrano, G., Mirats, J.J., Sarrate, R.: Leak localization in water networks. IEEE Control Systems Magazine **August**, 24–36 (2014)
24. Pudar, R.S., Liggett, J.A.: Leaks in pipe networks. Journal of Hydraulic Engineering **118**(7), 1031–1046 (1992)
25. Puust, R., Kapelan, Z., Savić, D.A., Koppel, T.: A review of methods for leakage management in pipe networks. Urban Water Journal **7**(1), 25–45 (2010)
26. Ragot, J., Maquin, D.: Fault measurement detection in an urban water supply network. Journal of Process Control **16**, 887 (2006)

27. Sarrate, R., Blesa, J., Nejjari, F., Quevedo, J.: Sensor placement for leak detection and location in water distribution networks. Water Science and Technology: Water Supply **14**(5), 795–803 (2014)
28. Savić, D.A., Kapelan, Z., Jonkergouw, P.: Quo vadis water distribution model calibration? Urban Water Journal **6**(1), 3–22 (2009)
29. Soldevila, A., Blesa, J., Tornil-Sin, S., Duviella, E., Fernandez-Canti, R.M., Puig, V.: Leak localization in water distribution networks using a mixed model-based/data-driven approach. Control Engineering Practice **55**, 162–173 (2016)
30. Stephens, M.A.: EDF statistics for goodness of fit and some comparisons. Journal of the American statistical Association **69**(347), 730–737 (1974)
31. Thornton, J., Lambert, A.: Progress in practical prediction of pressure: leakage, pressure: Burst frequency and pressure: Consumption relationships. In: eakage 2005 Conference Proceedings. Halifax, Canada. (2005)
32. Todini, E., Pilati, S.: Computer applications in water supply: vol. 1—systems analysis and simulation. chap. A gradient algorithm for the analysis of pipe networks, pp. 1–20. Research Studies Press Ltd., Taunton, UK, UK (1988)
33. Wachla, D., Przystalka, P., Moczulski, W.: A method of leakage location in water distribution networks using artificial neuro-fuzzy system. IFAC-PapersOnLine **48**(21), 1216 – 1223 (2015)
34. Wu, Z.Y., Sage, P.: Water loss detection via genetic algorithm optimization-based model calibration. In: Systems Analysis Symposium, pp. 1–11. ASCE (2006)
35. Yang, J., Wen, Y., Li, P.: Leak location using blind system identification in water distribution pipeline. Journal of Sound and Vibration (310), 134–148 (2008)