# Leak Localization in Water Distribution Networks using a Mixed Model-Based/Data-Driven Approach

Adrià Soldevila[a], Sebastian Tornil-Sin[a,b], Joaquim Blesa[a,b], Eric Duviella[c,d],
Rosa M. Fernandez-Canti[a], Vicenç Puig[a,b,*]

[a] *Research Center for Supervision, Safety and Automatic Control (CS2AC).*
*Rambla Sant Nebridi, s/n, 08022 Terrassa (Spain).*
[b] *Institut de Robòtica i Informàtica Industrial (CSIC-UPC).*
*Carrer Llorens Artigas, 4-6, 08028 Barcelona (Spain).*
[c] *Mines-Telecom Institute, Mines Douai, France*
[d] *Université de Lille, Lille, France*

## Abstract

This paper proposes a new method for leak localization in water distribution networks (WDNs). In a first stage, residuals are obtained by comparing 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. The proposed method is tested both using synthetic and experimental data with real WDNs of different sizes. The comparison with the current existing approaches shows a performance improvement.

*Keywords:* Water distribution networks, leak localization, fault diagnosis, model-based methods, data-driven methods, Barcelona WDN.

## 1. Introduction

Water leaks in a water distribution network (WDN) can cause significant economic losses in fluid transportation leading to increase reparation costs that finally generate an extra cost for the final consumer. In many WDN, losses due

[*] Corresponding author.
*Email address:* vicenc.puig@upc.edu (Vicenç Puig)

to leaks are estimated to account up to 30 % of the total amount of extracted water [1]. This is a very important amount in a world struggling to satisfy water demands of a growing population [2, 3, 4, 5].

The traditional approach to leakage control is a passive one, whereby the leak is repaired only when it becomes visible. Recently developed acoustic instruments [6] 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 [7, 1], and to maintain a permanent leakage control-system: leakages in fact increase the flow and decrease the pressure measurements at the DMA entrance. Various empirical studies [8, 9] propose mathematical models to 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 flow is minimum. This typically occurs at night, when customers' demand is low and the leakage component is at its largest percentage over the flow [1]. Therefore, practitioners monitor the DMA or groups of DMAs for detecting (and then repairing) leakages by analyzing the minimum night flow, and also employ techniques to estimate the leakage level [1]. However, leakage detection may not be easy, because of unpredictable variations in consumer demands and measurement noise, as well as long-term trends and seasonal effects.

Several works have been published dealing with leak location methods for WDN (see [10] and references therein). For example, in [11], a review of transient-based leak detection methods is offered as a summary of current and past work. In [12], a method is proposed to identify leaks using blind spots based on previously leak detection that uses the analysis of acoustic and vibrations signals [13], and models of buried pipelines to predict wave velocities [14]. More recently, [15] have developed a method to locate leaks using Support Vector Machines (SVM) that analyzes data obtained by a set of pressure control sensors of a pipeline network to locate and calculate the size of the leak. Another set of methods is based on the inverse transient analysis [16, 17]. The main idea of this methodology is to analyze the pressure data collected during the occurrence of transitory events by means of the minimization of the difference between the observed and the calculated parameters. In [18, 19], 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.

Model-based leak detection and isolation techniques have also been studied starting with the seminal paper of Pudar [20] 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 [21]. The problem of leak localization in WDNs can be addressed as a particular case of the general problem known in the literature as the problem of Fault Detection and Isolation (FDI) in dynamic systems [22]. However, the model of a DMA leads to a non-explicit model that can only be solved using numerical methods and limiting the applicability of most of the current FDI approaches that make an explicit use of the model. Moreover, there exist a high coupling of residuals and leaks plus a reduced number of sensors that as a result they complicate the isolation task. For this reason specific fault diagnosis methods for leak localization should be developed. A first contribution in this line can be found in [23, 24] where a model-based method that relies on pressure measurements and leak sensitivity analysis is proposed. This methodology consists in computing on-line residuals, i.e. differences between the measurements and their estimations obtained using the hydraulic network model, and checking them against thresholds that take into account the modeling uncertainty and the noise. When some of the residuals violate their threshold, the residuals are matched against the leak sensitivity matrix in order to discover which of the possible leaks is present. Although this approach has good efficiency under ideal conditions, its performance decreases due to the nodal demand uncertainty and noise in the measurements. This methodology has been improved in [25] where an analysis along a time horizon has been taken into account and a comparison of several leak isolation methods is presented. It must be noticed that in cases where flow measurements are available, leaks could be detected more 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 using fuzzy analysis of the residuals. This method finds the residuals between the flow measurements and their estimation using a model without leaks. However, although the use of flow measurements is feasible in large water transport networks, this is not the case in water distribution networks where there is a dense mesh of pipes with only flow measurements at the entrance of each District Metering Area (DMA). In this situation, water companies consider as a feasible approach the possibility to installing some pressure sensors inside the DMAs, because they are cheaper and easier to install and maintain.

In this paper, a new approach for leak localization in WDNs is presented. This methodology is used once the leak has been detected by means of the analysis of the nightly water demands of the DMA that is used for detecting and estimating the leakage level [1], and after the application of the sensor validation and reconstruction described in [27]. The approach combines the use of pressure models and classifiers. Following a model-based methodology successfully tested in [23] and [24], a pressure model of the considered WDN is used in a first stage to compute residuals, i.e. differences between the measured (sensors) and estimated (model) values of the water pressure in nodes of the network, 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 in extensive simulations of the network. These simulations consider 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.

The rest of the paper is organized as follows. Section 2 presents the background on model-based fault leak localization methods based on sensitivity analysis and highlights their limitations. Section 3 presents in detail the proposed method. Section 4 details the application of the method to three WDNs of different sizes and provides a comparison with other well-established approaches. Finally, Section 5 draws the main conclusions of the work and introduces some potential extensions.

## 2. Background

### *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 using the hydraulic network model. Usually, this methodology is used for locating leaks in 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 burst and the leak can be seen in 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 using the DMA network mathematical model [24]. This comparison uses the residual vector, $\mathbf{r} \in \mathbb{R}^{n_s}$ that is determined by the difference between the measured pressure at inner nodes where sensors are installed

$$\mathbf{r}(t) = \mathbf{p}(t) - \hat{\mathbf{p}}_{\mathbf{o}}(t) \tag{1}$$

where $\mathbf{p} \in \mathbb{R}^{n_s}$, and the estimated pressure at these nodes obtained using the network model considering a leak-free scenario, $\hat{\mathbf{p}}_{\mathbf{o}} \in \mathbb{R}^{n_s}$

The size of the residual vector $\mathbf{r}$, $n_s$, depends on the number of inner pressure sensors installed in the DMA. In recent years, some optimal pressure sensor placement algorithms have been developed to determine which pressure sensors have to be installed inside the DMA such that with minimum economical costs (number of sensors), a suitable performance regarding leak localization is guaranteed, see [23], [28], [29] among others.

The number of potential leaks, $\mathbf{f} \in \mathbb{R}^{n_n}$, is considered to be equal to the number of DMA nodes $n_n$, since from the modeling point of view, as proposed in [23] and [24], leaks are assumed to be in these locations.

### *2.2. Limitations of sensitivity analysis approaches*

Most of the model-based leak localization approaches rely on the sensitivity-to-leak analysis [23, 24] 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 (DMA network nodes, $n_n$)). Then, leak isolation is based on matching the residual vector (1) against the columns of the sensitivity matrix using some metrics (see [25] for details). The leak sensitivity matrix can be mathematically formalized as follows

$$\Omega = \begin{pmatrix} \frac{\partial p_1}{\partial f_1} & \cdots & \frac{\partial p_1}{\partial f_{n_n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{ns}}{\partial f_1} & \cdots & \frac{\partial p_{ns}}{\partial f_{n_n}} \end{pmatrix} \tag{2}$$

where each element $\Omega_{ij}$ measures the effect of the leak $f_j$ in the pressure $p_i$ of the node where the inner pressure sensor $i$ is located. However, in practice, it is extremely difficult to calculate $\Omega$ analytically in a real network because a water network is a large scale problem described by a multivariable non-linear system of equations which may also be non-explicit. Thereby, the sensitivity matrix is generated by simulation of the network model approximating the sensitivity $\Omega_{ij}$ by

$$\Omega_{ij} = \frac{\hat{p}_{if_j} - \hat{p}_{i0}}{f_j} \tag{3}$$

where $\hat{p}_{if_j}$ is the predicted pressure in the node where the pressure sensor $i$ is placed when a nominal leak $f_j$ is forced in node $j$ and $\hat{p}_{i0}$ is the predicted pressure associated with the sensor $i$ under a scenario free of leaks [23]. Then, repeating this process for all $n_n$ potential faults the approximation of the sensitivity matrix is obtained.

Another difficulty of the leak sensitivity approach is that the practical evaluation of (3) depends on the nominal leak $f_j$ [30, 31]. 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 [32] since it is not measured but estimated using historical records of water consumption and using the aggregated DMA consumption pattern. These uncertainties will lead to worsen the leak localization results obtained using the the leak sensitivity approach. The approach proposed in this paper aims to overcome these difficulties.

## 3. Proposed method

### *3.1. A mixed model-based/data-driven FDI approach*

Two main types of approaches used to face the FDI problem are the model-based and the data-driven approaches, both with their advantages and drawbacks [33]. Model-based FDI methods rely on the generation and evaluation of residuals. A residual is a computable (on-line) expression derived from the system model that evaluates close to zero in absence of faults and that deviates from zero in response to some of the faults. The design (off-line) of a model-based FDI system typically relies on the manipulation of a system model that includes the effect of faults, with the aim of obtaining a set of residuals that are sensitive to the faults in such a way they can be distinguished. In particular, some of the methods proposed in the literature try to obtain a set of residuals with each one of them sensitive to a different subset of the considered faults (structured residuals), since then fault isolation is straightforward under a Boolean framework. Other methods try to obtain residual vectors that follow different directions according to the presence of different faults (directional residuals) and fault isolation is solved under a geometric framework. The main drawback in the application of model-based methods is that a deep knowledge of the system operation (physical equations) and an important modeling effort (estimating the model parameters) are required.

On the other hand, data-driven approaches do not need a priori knowledge (model) about the monitored system since they directly use experimental data. Assuming that the faults affect some observable system variables and that historic data is available, the fault isolation problem can be formulated as a pattern recognition problem. If the available data is labeled, i.e. each observation of the symptoms is known to be obtained under a given faulty situation, then fault isolation can be formulated as a classification problem. If not, fault isolation can still be formulated as a clustering problem. The main drawback of using a data-driven approach is that the designed fault isolation system will only be able to deal with faults that have been previously experimented by the monitored system. In practical applications, the available data can be limited to a subset of the possible faults that the system can experiment and that are of interest. An additional drawback when using a classification approach is that it may be difficult or even not possible to guarantee a correct labeling of the available data.

The method for leak localization proposed in this paper can be considered as a mixed model-based/data-driven FDI method. The method uses a model of the monitored water network to generate pressure residuals. The model is implemented in the Epanet [34] hydraulic simulator[1] which is the standard simulation tool in water distribution networks and it is the software used for the company responsible of the Barcelona WDN management. These are the reasons why EPANET has been chosen as hydraulic simulator in this paper. The obtained residuals are raw residuals that do not facilitate a straightforward fault isolation under a Boolean or geometric framework. The reasons for this is that all the leaks affect to some extent to all the residuals and that there is a high level of uncertainty: the leak size is unknown, the nodal demands are also

[1] There exists other hydraulic simulators, for a review see [35].

unknown and have to be estimated and, finally, measurements are affected by noise. Due to this, it is proposed to use these raw residuals to feed a classifier. It must be noticed that in this context the use of the classifier do not suffer the problems associated to purely data-driven methods, since availability of the simulator allows to generate complete data sets that cover all possible faults and that are perfectly labeled.

### *3.2. Basic architecture and operation*

The method for on-line leak localization proposed in this paper relies on the scheme depicted in Figure 1, based on computing pressure residuals and analyzing them by a classifier. Residuals are computed as differences between measurements provided by pressure sensors installed inside the DMA and estimations provided by a hydraulic model simulated under leak-free conditions. The hydraulic model is built using the Epanet hydraulic simulator by considering the DMA structure (pipes, nodes and valves) and network parameters (pipe coefficients) and it is assumed to be able to represent precisely the WDN behavior after the corresponding calibration process using real data. However, it must be noticed that the model is fed with estimated water demands (typically obtained by the total measured DMA demand $d_{WDN}$ and distributed at nodal level using historical consume records) in the nodes $(\hat{d}_1, \cdots, \hat{d}_{n_n})$ since 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). Hence, the residuals are not only sensitive to faults but also to differences between the real demands and their estimated values. Additionally, pressure measurements are subject to the effect of sensor noise $\boldsymbol{v}$ and this also affects the residuals. Taking all of 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, it must be noticed that the operation of the network is constrained by some boundary conditions (for example the position of internal valves, reservoir pressures and flows) that are known (measured) and that are taken into account in the simulation and that can also be used as inputs for the classifier.

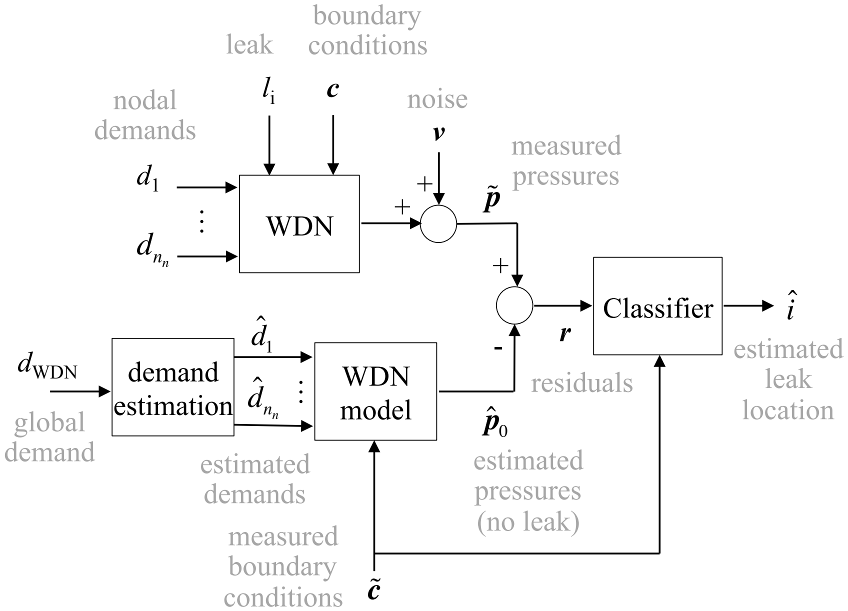


Figure 1: Leak localization scheme.

### 3.3. Methodology description: off-line training

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 paper considers an off-line design based on the following stages:

- Modelling - A model for the WDN is obtained, calibrated and implemented in Epanet. The model is basically built by taking into account the network structure and by applying flow balance conservation and pressure loss equations, see [23, 24] for details.

- Data generation - The model implemented in Epanet is extensively used to generate data in the residual space for each possible fault and for different operating and uncertainty conditions.

- Grouping - Nodes for which the effects of leaks in the residual space are similar are aggregated in groups of nodes before training the classifier.
- Classifier training and evaluation - The classifier is first trained to perform the classification process by using the training data, a subset of the initial data set, then it is applied to the validation data set in order to estimate its performance.

### *3.3.1. Data generation*

The data generation stage is critical since the availability of representative data is a necessary condition for obtaining a good classifier. Since the data that can be obtained from the real monitored WDN can be really limited, the way to obtain a complete training data set is by using the hydraulic simulator. Hence, training data, and also validation data, is generated by applying the scheme depicted in Figure 2, 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.

The presented schemed is exploited in order to:

- Generate data for all possible leak locations, i.e. for all the different nodes in the WDN ($\bar{f}_i, \quad i = \{1, 2, ..., n_n\}$).
- For each possible leak location, generate data for different leak magnitudes inside a given range ($\bar{f}_i \in [f_i^-, f_i^+]$).
- Generate sequences of demands and boundary conditions 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 ($\bar{\nu}$).

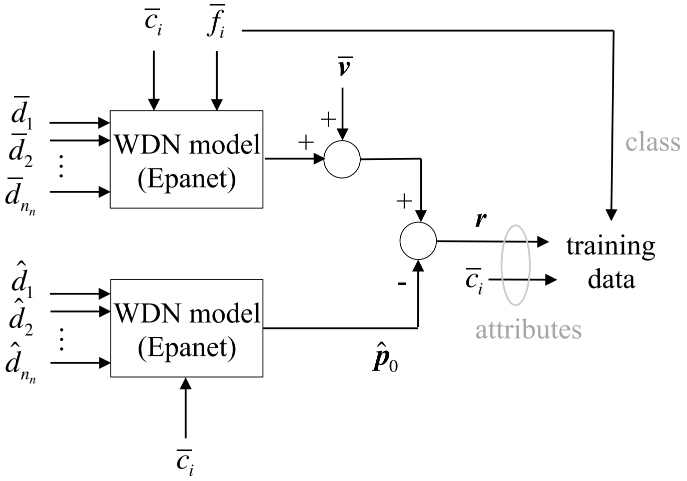


Figure 2: Data generation scheme.

*3.3.2. Node grouping*

If the described leak localization methodology is directly applied to a real WDN, it may be observed that the leaks in some nodes cannot be distinguished. This is mainly due to the present uncertainty, i.e. to the unknown value for the leak magnitude, to the differences between the real and the estimated water demands in the nodes and to measurement noises. But even in absence of uncertainty, the leaks in some nodes can be indistinguishable. This is for instance the case when some nodes are located in a same branch of the network and none of these nodes are equipped with a pressure sensor.

Since leaks in some nodes can not be distinguished because they present a very similar leak signature, a node grouping prior to the classifier training is proposed. Nodes whose leaks present very similar effect in pressure sensors will be grouped in the same class creating a composed class. In particular, nodes $i$ and $j$ will be grouped in the same class if it is satisfied the following condition

$$\frac{1}{24}\sum_{t=1}^{24}\|\mathbf{r}_i^0(t)-\mathbf{r}_j^0(t)\| < \frac{\gamma\left\|\bar{\mathbf{r}}^0\right\|}{100} \tag{4}$$

where $\mathbf{r}_i^0(t)$ and $\mathbf{r}_j^0(t)$ are the nominal pressure residuals at time $t$ obtained after introducing a leak in nodes $i$ and $j$ with the same nominal conditions ($\tilde{\boldsymbol{v}} = 0$, $\tilde{\mathbf{d}}(t) = \hat{\mathbf{d}}(t)$ and $f_i = f_j = f^0$), and $\bar{\mathbf{r}}^0$ is the daily average pressure residual computed as

$$\left\|\bar{\mathbf{r}}^0\right\| = \frac{\sum_{i=1}^{n_n}\sum_{t=1}^{24}\|\mathbf{r}_i^0(t)\|}{24n_n} \tag{5}$$

and $\gamma$ is a predefined threshold.

### *3.3.3. Classifier training and evaluation*

After the grouping process, the data is divided into training and validation data sets that will be used in the associated stages. The training stage is a learning from examples 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. This generalization ability of the obtained classifier is checked in the validation stage, in which the classifier is applied to the validation data set and performance indexes are computed.

The details of the training stage are particular of the type of classifier used. The results presented in this paper have been obtained by using the well known $k$-Nearest Neighbor ($k$-NN) classifier [36]. This classifier is said to be a type of *lazy* classifier since the training stage is limited to the recording of the training data set and all the required computations are deferred until the classifier is used to classify new data instances (see details in Section 3.4.1).

The evaluation of classifiers normally relies on the use of the *confusion matrix* $\Gamma$, that summarizes the results obtained when the classifier is applied to the validation data set. Applied to the leak localization problem and using the associated terminology, the confusion matrix is a square matrix with as many rows and columns as nodes in the network (potential leak locations), where each coefficient $\Gamma_{ij}$ 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 (in general, to fault isolation).

Table 1: Confusion matrix $\Gamma$

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

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

$$Ac = \frac{\sum_{i=1}^{n_n} \Gamma_{ii}}{\sum_{i=1}^{n_n} \sum_{j=1}^{n_n} \Gamma_{ij}} \tag{6}$$

Accordingly to the previous paragraphs, the confusion matrix is normally used to compute a set of performance indexes about the behavior of the classifier. In this paper, as it will be presented later, the confusion matrix is also used to complement the classifier in order to obtain better leak localization results (see details in Section 3.4.2).

### *3.4. Methodology description: on-line leak location*

Once the classifier has been validated, it can be used on-line to localize leaks. According to Section 3.2, the classifier can be directly used to estimate the presence of leaks by applying it to the instantaneous values for the computed residuals. However, this strategy may provide limited results in presence of a high level of uncertainty. This suggests the use of a temporal reasoning that takes into account not only the instantaneous values for the residuals but all the values in a time horizon. The basic operation of the used $k$-NN classifier and the details of the proposed temporal reasoning strategy are presented below.

### *3.4.1. The k-NN classifier*

One of the well accepted and established methods for classification is the $k$-Nearest Neighbors ($k$-NN) algorithm [36] which is available in most of the numerical packages (e.g. Matlab, R, etc.). Its basic version works as follows. When a new data realization has to be classified, the distances [2] to all the instances in the training data set are 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 as the associated class for the new data instance. 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 are associated to training instances in that class.

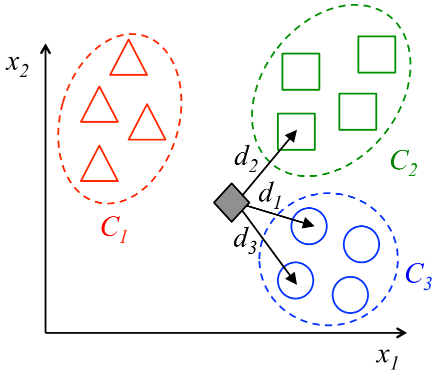


Figure 3: The $k$-NN algorithm.

The use of a value for $k$ bigger than one improves the robustness against outliers (with $k = 1$ the class of the nearest neighbor is selected, which seems a good choice, but the obtained classifier is really sensitive to outliers). On

[2] Typically, the Euclidean distance is used, but many other options are available.

the other hand, the value for $k$ must be smaller than the minimum number of instances associated to a single class inside the training data set.

### *3.4.2. Temporal reasoning*

If the uncertainty in the demands, 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 [37].

Under the framework proposed in this paper, a simple temporal reasoning can be based on taking into account the estimations provided by the classifier inside the time horizon and applying a 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 for the residuals in 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 being present in node $i$ when the classifier is indicating that the leak is in node $j$, according to the available information available for current time instant $t$. Then the sum of column vectors stored along the time horizon N $t-N+1,\cdots,t$ is computed. In the obtained vector, the position of the coefficient with highest value indicates the most probable fault according to the information provided by the data in the whole time horizon $t-N+1,\cdots,t$.

### *3.5. Summary*

For a better understanding of the whole proposed methodology, Figure 4 summarizes graphically its main steps. Both off-line design and on-line leak localization procedures described in previous sections are included. Moreover, the connections between both procedures are made clear: the hydraulic model (Epanet model), the classifier and the confusion matrix resulting from the modeling, training and evaluation modules of the off-line procedure are provided to the on-line procedure in order to perform the residual generation, classification and the temporal reasoning operations.

Note that in this figure a sensor validation/reconstruction and leak detection modules have been added. Although the implementation of these modules is out of the scope of the paper (as discussed in the introduction), they are necessary in the on-line procedure and they have been included in the figure for completeness. If the sensor validation/reconstruction module detects one sensor fault and it is not possible to reconstruct the measurement, the off-line design process, discarding the faulty sensor, is executed again. On the other hand, the leak detection module detects if a leak is present in the system and triggers the leak localization procedure proposed in this paper.

## 4. Case studies

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

For these DMAs, leaks are considered that could appear in any of the demand nodes. The known variables are the input pressures and flows of the networks (reservoir boundary conditions) and some pressures in inner nodes of the DMAs where sensors would be located (see [28, 29, 31] 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 in [32]. 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 three DMAs, leak localization results under different uncertainty

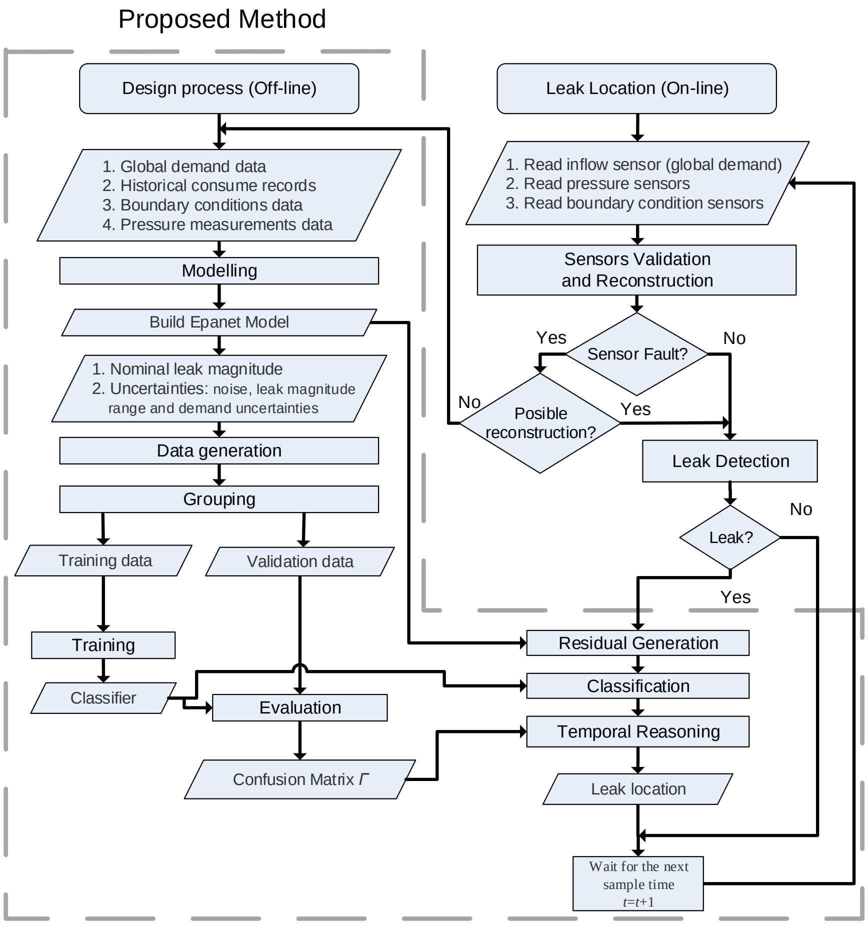


Figure 4: on-line and off-line schemes of the proposed approach

scenarios and obtained by using artificial (generated by simulation) data are presented and discussed. Moreover, for the last (and biggest) DMA the results of localizing a real leak are also presented.

### *4.1. Hanoi case study*

The Hanoi DMA network in Vietnam, depicted in Figure 5, consists of one reservoir, 34 pipes and 31 nodes. Two pressure sensors placed in internal nodes 14 and 30 have been considered (see [28] for more details).

The process of grouping leaks in composed classes (Section 3.3.2) has been carried out considering the satisfaction of condition (4), with $\gamma = 0.5$ and $f^0 = 50\,[\mathrm{l/s}]$. In result, two composed leak classes have been obtained. These two composed classes are depicted in Figure 5 and summarized in Table 2. On the other hand, the number of non-composed (single node) classes is 22. The composed classes obtained in this DMA show that using a reduced number of inner pressure sensors, some leaks could not be distinguished since they present an effect in the residuals very similar according to condition (4).

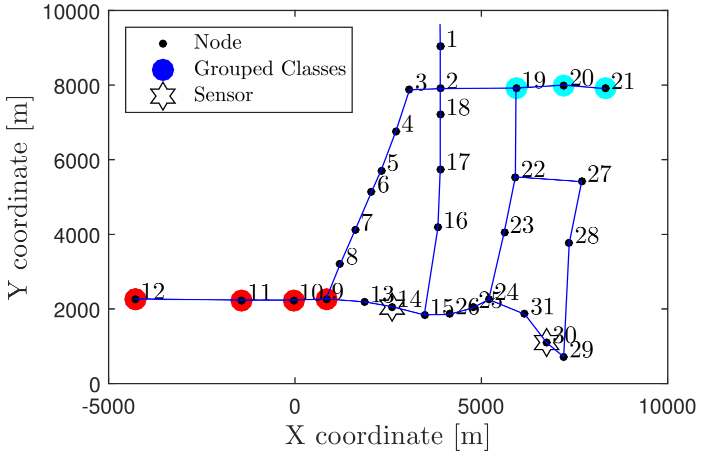


Figure 5: Hanoi topological network.

Table 2: Groups not separable in Hanoi Network.

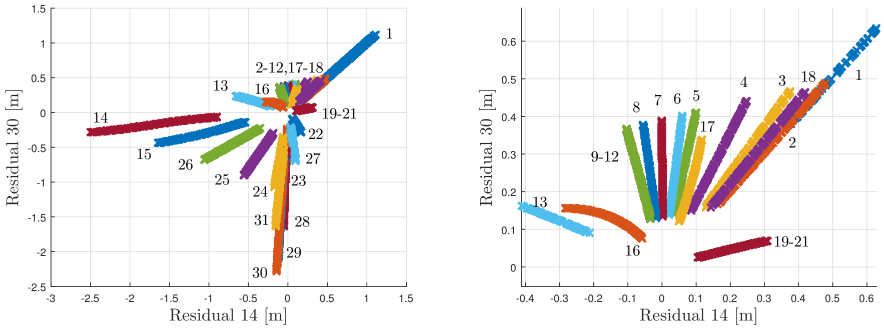
| Composed classes | | |
|---|---|---|
| Groups | Number of single classes | Single classes |
| Group 1 | 4 | 9, 10, 11, 12 |
| Group 2 | 3 | 19, 20, 21 |

In order to illustrate the performance of the proposed methodology under the effect of the different sources of uncertainty, three different studies have been carried out under the following associated conditions:

- The leak magnitude is considered to be unknown but inside the range between 25 and 75 [l/s] (0.84 and 2.51 % of the total amount of water demanded, which is 2991.1 [l/s]).
- Pressure measurements are affected by Gaussian noise with an amplitude of ± 12.5 % of the mean value of all pressure residuals.
- For each node, the instantaneous demand is unknown but inside the interval given by an uncertainty of ± 5 % around the nominal demand (estimated according to the node demand pattern and the monthly demand average).

For each study, a training data set and a validation data set have been generated by taking into account 200 and 50 samples (pressure residual vectors computed every hour) respectively for each possible leak location (i.e. for each node in the network). The residuals generated for training in the leak uncertainty study are depicted in Figures 6a and 6b, where different colors are used to identify the residuals obtained for different leak locations. Those residuals show the directional effect of the leak magnitude in the residuals, that is exploited by some leak localization approaches (see [25] for details). The residuals considered in the noise study are shown in Figure 7, where it can be seen a variation around the nominal residuals (leak size of 50 [l/s]) due to the presence of noise in pressure measurements. Finally, Figure 8 shows the training residuals in the demand uncertainty study, which present a similar effect as the noise study but with larger variation.

The results obtained by the proposed method in the three different studies have been compared with the ones obtained using the leak-sensitivity analysis with the angle metrics proposed in [25] and summarized in Section 2. For this purpose, the sensitivity matrix (2) has been computed using (3) considering nominal leak conditions in every demand node ($\tilde{\boldsymbol{v}} = 0$, $\tilde{\mathbf{d}} = \hat{\mathbf{d}}$ and $f_i = 50$ [l/s] $i = 1, ..., n_n$). The results obtained by using the two methods, in both cases



(a) General residual space.

(b) Residual space in detail.

Figure 6: Residual space with leak uncertainty in Hanoi network (different colors for different leak locations).

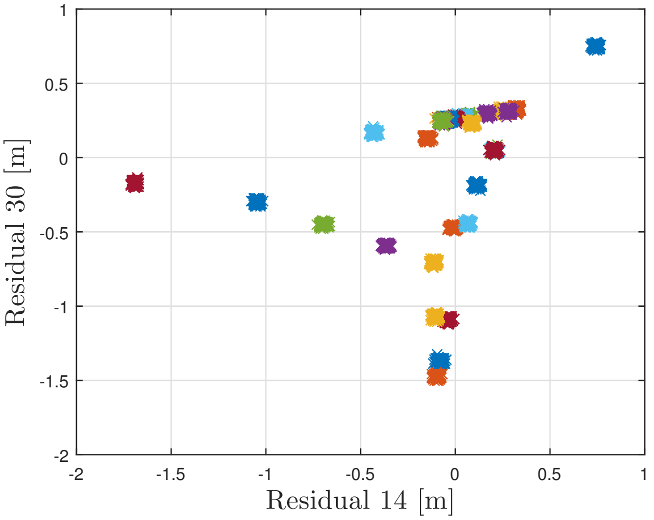


Figure 7: Residuals space with noise in Hanoi network (different colors for different leak locations).

considering only one sample in the leak location diagnosis ($N = 1$), are summarized in Table 3. The values presented in the table correspond to the overall accuracy $Ac$ defined in (6). As it can be seen, both methods provide good performances in the leak uncertainty case thanks to the linear directional variation of most of the residuals for this kind of uncertainty. In the case of noise in pres-

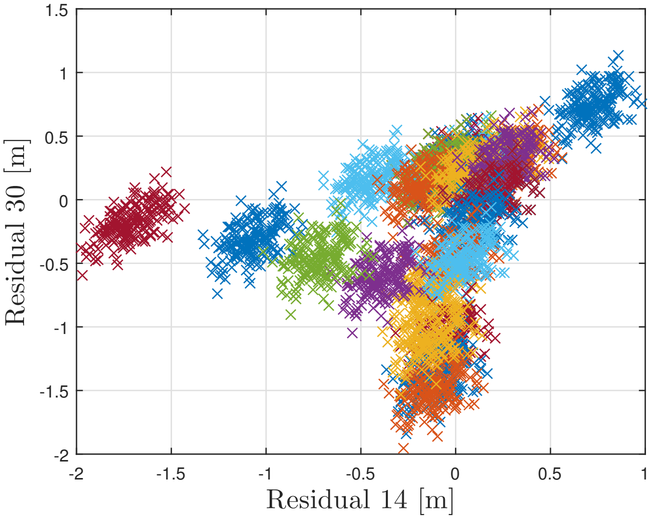


Figure 8: Residuals space with demand uncertainty in Hanoi network (different colors for different leak locations).

sure measurements, the method proposed in this paper performs significantly better. This is because the classifier can handle more efficiently the dispersion produced by noise in measurements. The same occurs when demand uncertainty is considered.

Table 3: Leak location results (overall accuracy $Ac$) for Hanoi network with $N = 1$.

| | Classifier | Angle |
|---|---|---|
| Leak uncertainty | 99.29 | 99.81 |
| Noise | 96.97 | 88.32 |
| Demand uncertainty | 43.48 | 30.26 |

The results obtained by the proposed method using a time horizon of $N = 24$ (that is, considering residual vector computed in the last 24 hours) and the confusion matrix, as described in Section 3.4.2, are summarized in Table 4. As it can be noticed, there is an improvement in performance achieved in all uncertainty cases, but especially in the demand uncertainty case.

Table 4: Leak location results (overall accuracy $Ac$) for Hanoi network with $N = 24$.

| | Classifier |
|---|---|
| Leak uncertainty | 100 |
| Noise | 100 |
| Demand uncertainty | 89.95 |

Finally, the effect of the horizon length $N$ in the performance of the proposed method considering demand uncertainty is shown in Figure 9. As it is expected, the accuracy increases with the time horizon length $N$. And it can be observed that it reaches a steady state value when $N$ is around twenty four (hours). This justifies the use of a time horizon corresponding to one day. This is in agreement with the results already presented in [25].

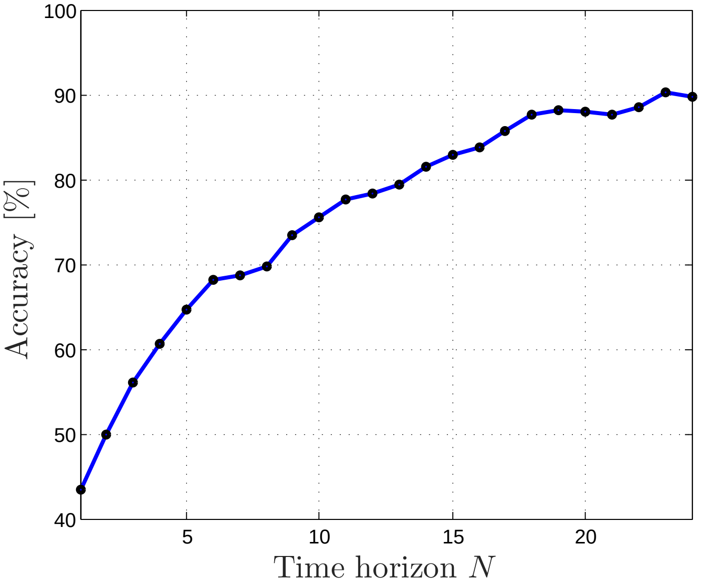


Figure 9: Classifier accuracy versus time horizon length $N$ in Hanoi network, considering demand uncertainty.

*4.2. Limassol case study*

The Limassol DMA network in Cyprus, shown in Figure 10, consists of one reservoir, 239 pipes and 197 demand nodes. Three pressure sensors that are placed following an optimal sensor placement described in [38] have been considered. In particular, the pressure sensors are placed in nodes 2, 146 and 152.

The process of grouping nodes has been carried out as in the Hanoi case study with a $\gamma = 0.5$ in (4). A number of 38 composed classes and 105 non-composed classes have been obtained. The composed classes are presented in Table 5. Composed classes 12, 18, 19 and 29 that group four or more nodes are highlighted in Figure 10.

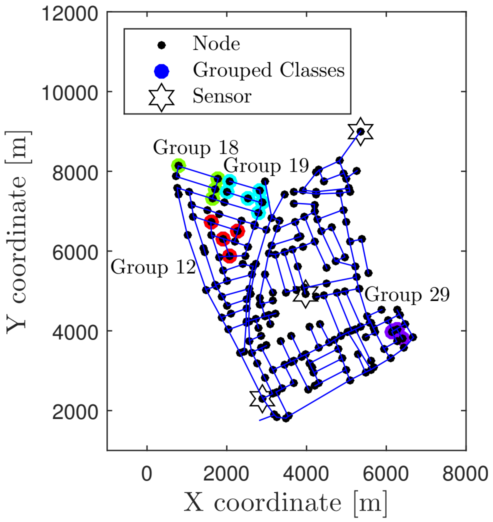


Figure 10: Limassol topological network.

For the Limassol network, a single study has been carried out by analyzing the effect of the combination of the three types of uncertainties already considered for the Hanoi network. In this study, the leak range has been considered between 3 and 5 [l/s] (0.61 and 1.02 % of the total amount of water demanded, which is 492.24 [l/s]), the noise amplitude in pressure measurements has been considered as the ± 5 % of the mean value of all residuals and the demand

Table 5: Groups non-separable in Limassol Network.

| Composed classes | | | |
|---|---|---|---|
| Groups | Single classes | Groups | Single classes |
| Group 1 | 9, 124 | Group 20 | 83, 84 |
| Group 2 | 18, 129 | Group 21 | 92, 94 |
| Group 3 | 20, 195 | Group 22 | 98, 130 |
| Group 4 | 21, 88 | Group 23 | 99, 105 |
| Group 5 | 22, 23 | Group 24 | 100, 102 |
| Group 6 | 29, 33 | Group 25 | 106, 108 |
| Group 7 | 36, 44 | Group 26 | 110, 111 |
| Group 8 | 37, 46 | Group 27 | 112, 113 |
| Group 9 | 38, 45 | Group 28 | 114, 115, 118 |
| Group 10 | 47, 48 | Group 29 | 116, 117, 132, 133 |
| Group 11 | 49, 50 | Group 30 | 136, 137 |
| Group 12 | 55, 59, 61, 62 | Group 31 | 142, 148 |
| Group 13 | 57, 64, 65 | Group 32 | 143, 144 |
| Group 14 | 58, 63 | Group 33 | 150, 191 |
| Group 15 | 60, 67 | Group 34 | 154, 155 |
| Group 16 | 69, 70, 186 | Group 35 | 156,158 |
| Group 17 | 71, 72, 77 | Group 36 | 157, 159 |
| Group 18 | 73, 74, 75, 76 | Group 37 | 168, 171, 173 |
| Group 19 | 78, 79, 80, 81, 82, 188 | Group 38 | 169, 170, 177 |

uncertainty has been considered as the ± 5 % of the nominal demand value. The sets of data for training and validation have the same size as in the Hanoi case study (200 and 50 samples, respectively).

As for the Hanoi network, the obtained results by applying the proposed method improve as the horizon length increases. The effect of the window length in the accuracy is shown in Figure 11. In particular, an 86.86 % of accuracy is reached when $N = 24$. On the other hand, Figure 12 shows the effect of the window length in the "Average Topological Distance", which is the average distance obtained from the minimum number of nodes between the node or group of nodes classified and the node or group of nodes where the leak belongs in any diagnosis and their unit is [nodes]. As it is expected, the average topological distance decreases with the increase of $N$, and a value around 0.2 is obtained for $N = 24$.

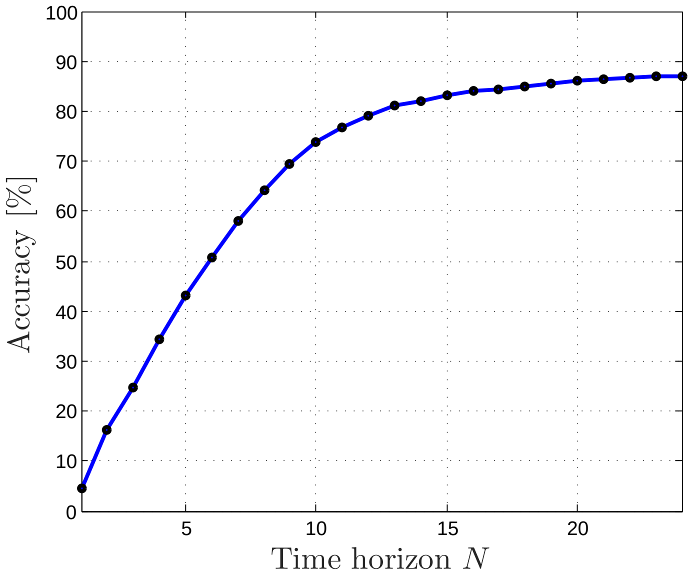


Figure 11: Accuracy versus time horizon length $N$ in Limassol network.

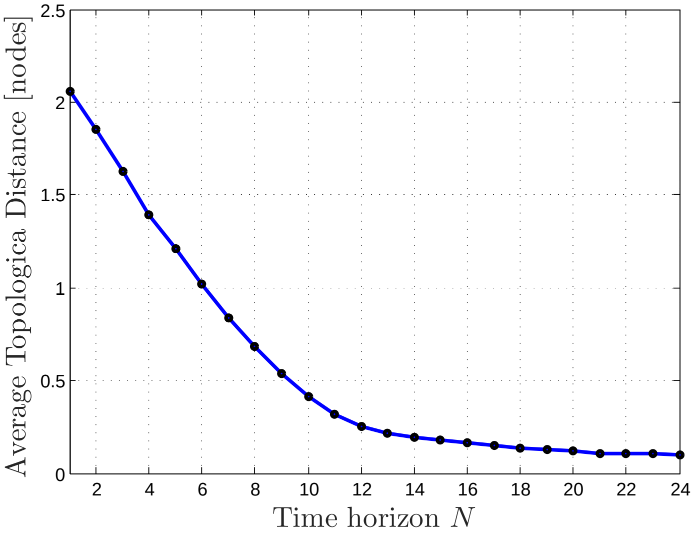


Figure 12: Average topological distance versus time horizon length $N$ in Limassol network.

### *4.3. Nova Icària case study*

The Nova Icària network, shown in Figure 13, is one of the DMA networks which form the Barcelona WDN. This network consist in 1520 nodes, 1646 pipes, two reservoirs and two valves, each one after the reservoirs with the aim of maintain 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 with previous network examples, some leak localization studies have been carried out in 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 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

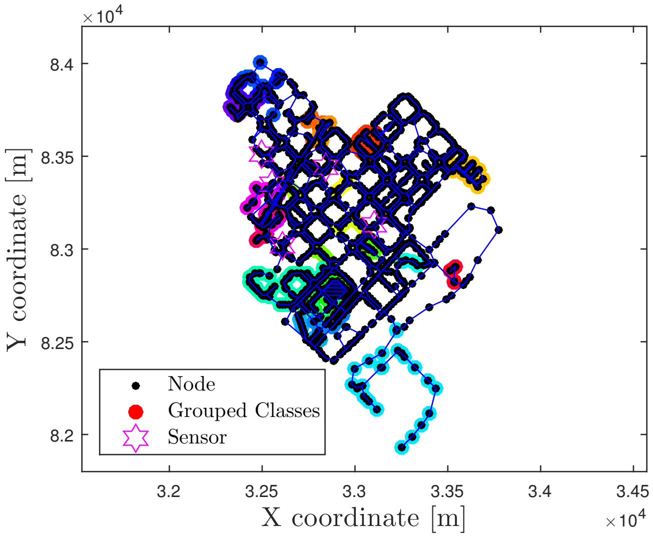


Figure 13: Nova Icària topological network.

see [24]. 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 $\mathbf{r}_i^0(t)$, $i = 1, \ldots, n_n$, $t = 1, \ldots, 24$ 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. These nominal hourly residuals have been used in the process of grouping nodes with a $\gamma = 0.5$ in (4). As a result of the grouping process, the 1520 nodes have been grouped in 1035 groups. The groups with five or more nodes are highlighted in Figure 13, where it is possible to see "dark zones" (mostly in branches) in the network where the leak hardly can be separated from a multitude of nodes given the placement of the five inner sensors.

A leak localization $k$-NN classifier (with $k = 3$) has been trained and validated. The inputs of the classifier are: the five pressure residuals, the flow that enters the DMA and the two set points of the valves. The data used in the training and validating processes are the 24 samples of nominal hourly residuals directly and adding the real uncertainty (120 samples): 96 samples for training and 48 for validation.

Given the size of this network and the limited number of sensor deployed inside the network, it is not realistic to expect a high degree of accuracy when the proposed diagnosis tries to locate the leak in the exact node, so the average topological distance is a more illustrative indicator to show how good the method performs. Taking into account the 1035 groups obtained in the grouping process and using a time horizon with $N = 24$, the average topological distance is around 7 (as it can be seen in Figure 14), a small value compared to the network size.

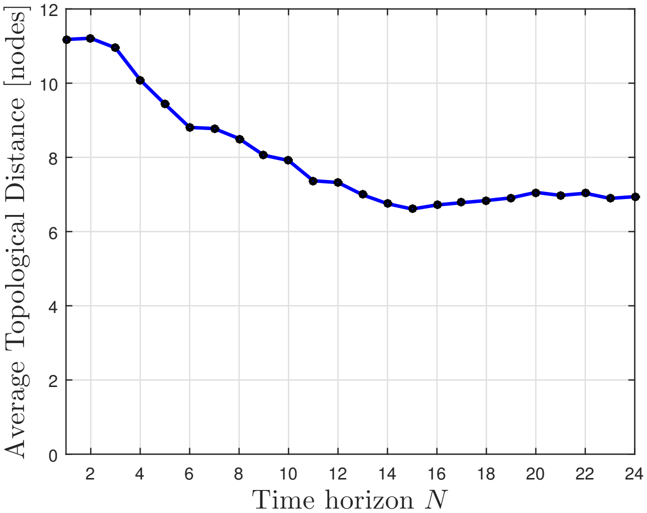


Figure 14: Average topological distance versus time horizon length $N$ in Nova Icària network.

Finally, the data of the real leak scenario has been applied to the trained classifier. Figure 15 shows the result of the proposed method after applying 24 hourly samples: the classifier indicates that the leak is in node 3 while the real leak is in node 996. The topological distance between these two nodes is 13 nodes, while the geographical linear distance is around 184 meters. For this real leak, the application of correlation method ([24]) provides as node candidate the node 1036 (this result is also shown in Figure 15), which is at a distance of 17 nodes and 222 meters of the real leak location.

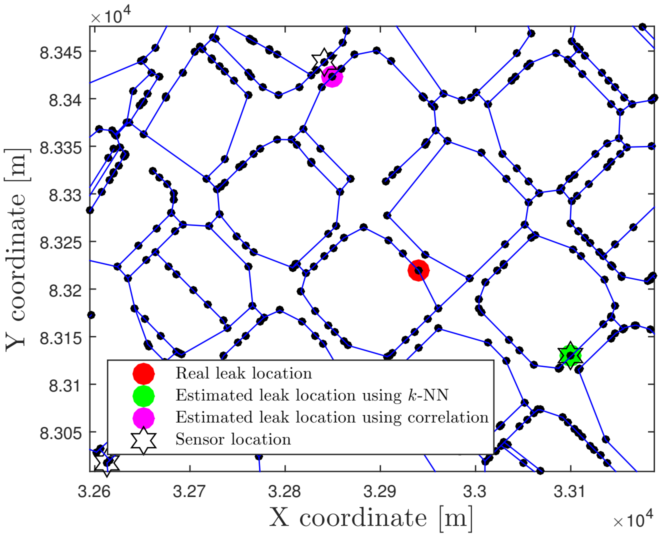


Figure 15: Comparison of different leak location methods in Nova Icària network.

## 5. Conclusion

This paper has proposed a new method for leak localization in WDNs that combines the use of pressure models with classifiers. Following a model-based methodology, 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 leak conditions. These simulations consider 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 proposed method has been compared with a previous leak localization method described in the literature through their application to three DMA case studies of different size and complexity obtaining satisfactory results.

The proposed approach has been developed assuming a single fault (leak). The extension to multiple faults is possible but it would require to train the classifier for the different possible multiple leak combinations considered. However, this could be very time consuming. Thus, as a further research more efficient methods to cope with the problem of multiple leaks will be adressed. Moreover, other type of classifiers will be considered that allow as e.g. to discover automatically from the structure of the network the leaks that present the same signature clustering them in the same group. Moreover, the extension of the proposed method to the case of faults evolving in time will be developed.

## 6. Acknowledgment

This work has been partially funded by the Spanish Ministry of Science and Technology through the Project ECOCIS (Ref. DPI2013-48243-C2-1-R) and Project HARCRICS (Ref. DPI2014-58104-R), and by EFFINET grant FP7-ICT-2012-318556 of the European Commission.

## References

[1] R. Puust, Z. Kapelan, D. A. Savic, T. Koppel, A review of methods for leakage management in pipe networks, Urban Water Journal 7 (1) (2010) 25–45.

[2] M. Valipour, A comprehensive study on irrigation management in Asia and Oceania, Archives of Agronomy and Soil Science (January 2015) (2014) 1–25.

[3] M. Valipour, Future of the area equipped for irrigation, Archives of Agronomy and Soil Science (July) (2014) 1–20.

[4] M. Valipour, Drainage, waterlogging, and salinity, Archives of Agronomy and Soil Science 0340 (March) (2014) 1–16.

[5] M. Valipour, Future of agricultural water management in Africa, Archives of Agronomy and Soil Science 0340 (January 2015) (2014) 1–21.

[6] Y. Khulief, A. Khalifa, R. Mansour, M. Habib, Acoustic detection of leaks in water pipelines using measurements inside pipe, Journal of Pipeline Systems Engineering and Practice 3 (2) (2012) 47–54.

[7] M. F. Lambert, A. R. Simpson, J. P. Vítkovský, X. J. Wang, P. J. Lee, A review of leading-edge leak detection techniques for water distribution systems, in: 20th AWA Convention, Perth, Australia, 2003.

[8] A. Lambert, 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.

[9] J. Thornton, A. Lambert, Progress in practical prediction of pressure: leakage, pressure: burst frequency and pressure: consumption relationships, in: eakage 2005 Conference Proceedings. Halifax, Canada., 2005.

[10] Z. Y. Wu, P. Sage, Water loss detection via genetic algorithm optimization-based model calibration, in: Systems Analysis Symposium, ASCE, 2006, pp. 1–11.

[11] A. F. Colombo, P. Lee, B. W. Karney, A selective literature review of transient-based leak detection methods, Journal of Hydro-environment Research (2009) 212–227.

[12] J. Yang, Y. Wen, P. Li, Leak location using blind system identification in water distribution pipeline, Journal of Sound and Vibration (310) (2008) 134–148.

[13] H. V. Fuchs, R. Riehle, Ten years of experience with leak detection by acoustic signal analysis, Applied Acoustics (33) (1991) 1–19.

[14] J. M. Muggleton, M. J. Brennan, R. J. Pinnington, Wavenumber prediction of waves in buried pipes for water leak detection, Journal of Sound and Vibration (249) (2002) 939–954.

[15] J. Mashford, D. D. Silva, D. Marney, S. Burn, 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, 2009, pp. 534–539.

[16] D. Covas, H. Ramos, Hydraulic transients used for leak detection in water distribution systems, in: 4th International Conference on Water Pipeline Systems, BHR Group, 2001, pp. 227–241.

[17] A. Kepler, D. Covas, L. Reis, Leak detection by inverse transient analysis in an experimental *pvc* pipe system, Journal of Hydroinformatics 13 (2) (2011) 153–166.

[18] M. Ferrante, B. Brunone, Pipe system diagnosis and leak detection by unsteady-state tests. 1. harmonic analysis, Advances in Water Resources 26 (1) (2003) 95–105.

[19] M. Ferrante, B. Brunone, Pipe system diagnosis and leak detection by unsteady-state tests. 2. wavelet analysis, Advances in Water Resources 26 (1) (2003) 107–116.

[20] R. S. Pudar, J. A. Liggett, Leaks in pipe networks, Journal of Hydraulic Engineering 118 (7) (1992) 1031–1046.

[21] D. Savic, Z. Kapelan, P. Jonkergouw, Quo vadis water distribution model calibration?, Urban Water Journal 6 (1) (2009) 3–22.

[22] M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, Diagnosis and Fault-tolerant Control, pringer-Verlag, Berlin/Heidelberg., 2006.

[23] R. Pérez, V. Puig, J. Pascual, J. Quevedo, E. Landeros, A. Peralta, Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks, Control Engineering Practice 19 (10) (2011) 1157 – 1167.

[24] R. Pérez, G. Sanz, V. Puig, J. Quevedo, F. Nejjari, J. Meseguer, G. Cembrano, J. J. Mirats, R. Sarrate, Leak localization in water networks, IEEE Control Systems Magazine August (2014) 24–36.

[25] M. V. Casillas, L. Garza-Castañón, V. Puig, 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, Elsevier, 2012, pp. 570–575.

[26] J. Ragot, D. Maquin, Fault measurement detection in an urban water supply network, Journal of Process Control 16 (2006) 887.

[27] M. A. Cugueró-Escofet, D. García, J. Quevedo, V. Puig, S. Espin, J. Roquet, A methodology and a software tool for sensor data validation/reconstruction: Application to the catalonia regional water network, Control Engineering Practice 49 (2016) 159 – 172.

[28] M. V. Casillas, V. Puig, L. E. Garza-Castañón, A. Rosich, Optimal sensor placement for leak location in water distribution networks using genetic algorithms, Sensors 13 (11) (2013) 14984–15005.

[29] R. Sarrate, J. Blesa, F. Nejjari, J. Quevedo, Sensor placement for leak detection and location in water distribution networks, Water Science and Technology: Water Supply 14 (5) (2014) 795–803.

[30] J. Blesa, V. Puig, J. Saludes, 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) (2012) 1890–1912.

[31] J. Blesa, F. Nejjari, R. Sarrate, Robust sensor placement for leak location: analysis and design, Journal of Hydroinformatics 18 (1) (2016) 136 – 148.

[32] P. Cugueró-Escofet, J. Blesa, R. Pérez, M. A. Cugueró-Escofet, G. Sanz, Assessment of a leak localization algorithm in water networks under demand uncertainty, IFAC Proceedings Volumes (IFAC-PapersOnline) 48 (21) (2015) 226–231.

[33] R. Isermann, Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance, Springer, 2006.

[34] L. A. Rossman, EPANET 2: users manual, Cincinnati US Environmental Protection Agency National Risk Management Research Laboratory 38.

[35] R. Schmid, Review of modelling software for piped distribution networks, English (2002) 1–18.

[36] E. Alpaydin, Introduction to Machine Learning, MIT Press, 2010.

[37] M. V. Casillas, L. Garza-Castañón, V. Puig, Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities, Journal of Hydroinformatics.

[38] M. V. Casillas, L. E. Garza-Castañón, V. Puig, A. Vargas-Martinez, Leak Signature Space: An Original Representation for Robust Leak Location in Water Distribution Networks, Water 7 (3) (2015) 1129–1148.