Leak Localization in Water Distribution Networks using Model-Based Bayesian Reasoning

Adrià Soldevila¹, Rosa M. Fernandez-Canti¹, Joaquim Blesa¹,², Sebastian Tornil-Sin¹,², and Vicenç Puig¹,²

Abstract—This paper presents a new method for leak localization in Water Distribution Networks that uses a model-based approach combined with Bayesian reasoning. Probability density functions in model-based pressure residuals are calibrated off-line for all the possible leak scenarios by using a hydraulic simulator, being leak size uncertainty, demand uncertainty and sensor noise considered. A Bayesian reasoning is applied online to the available residuals to determine the location of leaks present in the Water Distribution Network. A time horizon method combined with the Bayesian reasoning is also proposed to improve the accuracy of the leak localization method. The Hanoi District Metered Area case study is used to illustrate the performance of the proposed approach.

Keywords: Water distribution networks, leak localization, fault diagnosis, Bayesian reasoning, model-based.

I. 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 WDNs, losses due to leaks are estimated to account up to 30% of the total amount of extracted water. This is a very important amount in a world struggling to satisfy water demands of a growing population.

Several works have been published dealing with leak detection and isolation (localization) methods for WDNs (see [1] and references therein). For example, in [2], a review of transient-based leak detection methods is offered as a summary of current and past work. In [3], a method is proposed to identify leaks using blind spots based on previously leak detection that uses the analysis of acoustic and vibrations signals [4], and models of buried pipelines to predict wave velocities [5]. More recently, in [6], it has been 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. The use of classifiers in leak localization has been proposed in [7] and [8]. Another set of methods is based on the inverse transient analysis [9], [10]. 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 [11], [12], 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 [13] which formulates the leak detection and localization problem as a least-squares parameter estimation problem. However, the parameter estimation of water network models is not an easy task [14]. The difficulty lies in the non-linear nature of water network model and the few measurements usually available with respect to the large number of parameters to be estimated that leads to an underdetermined problem. Alternatively, in [15], [16], a model-based method that relies on pressure measurements and leak sensitivity analysis is proposed. This methodology consists in analyzing the residuals (difference between the measurements and their estimation using the hydraulic network model) online regarding a given threshold that takes 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 [17] 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 the 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 [18] 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 viable 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 of installing some pressure sensors inside the DMAs, because they are cheaper and easy to install and maintain.

In this paper, a new method for leak localization in WDNs that uses a model-based approach combined with Bayesian reasoning is proposed. Probability Density Functions (PDFs) in model-based pressure residuals are calibrated off-line for all the possible leak scenarios by using a hydraulic simulator.
and being leak size uncertainty, demand uncertainty and sensor noise considered. A Bayesian reasoning is applied on-line to the available residuals to determine the location of leaks present in the WDN. A time horizon method combined with the Bayesian reasoning is also proposed to improve the accuracy of the leak localization method. The Hanoi DMA case study is used to illustrate the performance of the proposed approach.

The paper is organized as follows. Section II presents in detail the proposed method while Section III describes the required calibration processes before its application to a WDN. Section IV details the application of the method to the Hanoi DMA case study and provides a comparison with another well-established approach. Finally, Section V draws the main conclusions of the work.

II. METHODOLOGY

A. Methodology overview

Following the approach proposed in previous works [15], [16], this paper proposes a method for leak localization based on the generation and analysis of pressure residuals. The contribution relies on the use of a Bayesian reasoning to implement such analysis.

1) Architecture: The on-line operation of the proposed method uses the classical two-stage model-based Fault Detection and Isolation (FDI) architecture based on generating and evaluating residuals. In particular, the leak localization method proposed in the paper considers on the scheme depicted in Fig. 1, based on computing pressure residuals and processing them by using a Bayesian reasoning procedure. Residuals are computed as differences between the measurements provided by pressure sensors installed inside the WDN and the estimations provided by a hydraulic model simulated under leak-free conditions. Residuals are sensitive to leaks since leaks change the real pressures and hence the measured values, whereas the corresponding reference estimated values remain unchanged. The aim of the Bayesian reasoning is to determine the most likelihood leak that explains the mismatches between the measurements and the estimations. It is assumed that just one leak can be affecting the network in a given time instant (single-fault assumption). The extension to multiple leaks will be addressed in future research.

The hydraulic model is built using a hydraulic simulator Epanet 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 analyzing historical billing records) in the nodes \( \hat{d}_1, \cdots, \hat{d}_n \) since in practice nodal demands \( d_1, \cdots, d_n \) 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 \( v \) that affects the residuals too. Taking all of these effects into account, the Bayesian reasoning 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.

2) Design: As with any other model-based FDI method, the method proposed in this paper will operate on-line but it requires a previous off-line design. In traditional FDI approaches, the design typically relies on the manipulation of the model (combination and/or projection of equations) with the aim of obtaining enhanced residuals, i.e. residuals that can facilitate isolation under a Boolean (structured residuals) or geometric (directional residuals) framework. Unfortunately, this type of procedure can not be applied to the localization of faults in WDNs since the model is a set of non-linear implicit equations that can not be easily manipulated. Alternatively, what can be done is generating synthetic data in the residual space by executing extensive simulations of the model under the different faulty conditions and trying to analyze directly the obtained raw residuals. In a previous work [7], this was done by training and using a classical \( k\)-Nearest Neighbors (\( k\)-NN) classifier. In this work, a Bayesian approach is proposed.

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 [15], [16] for details.
- **Data generation** - The model is extensively used to generate data in the residual space for each possible leak. The residuals correspond to differences between the pressures obtained considering fault scenarios and the ones obtained in the fault-free scenario.
- **Calibration** - The data associated to each fault is used to calibrate a joint probability density function that models the effect of each fault in the residual space. The calibration procedure is detailed in Section III of the paper.

![Fig. 1. Leak localization scheme.](image-url)
The data generation stage is critical since the generated data has to be complete to allow obtaining representative probability density functions and the maximum possible degree of isolability. Data is generated by applying the scheme presented in Fig. 2, that is similar to the one presented in Fig. 1 but substituting the real WDN by its Epanet model.

![Data generation scheme.](image)

The proposed scheme is exploited in order to:

- Generate data for all possible leak locations, i.e. for all the different nodes in the WDN.
- For each possible leak location, generate data for different leak magnitudes inside a given range.
- Generate data from sequences of demands that correspond to an estimated daily evolution of the demand in each node. These sequences are determined by taking into account the type of consumers that configure each nodal demand and the daily average demand.
- Take into account the uncertainty in the demands, i.e. the differences between the real demands and their estimated values. This is done by working with simulated real demands \(d_1, d_2, \ldots, d_n\) that differ from the simulated estimations \(\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_n\).
- Take into account the measurement noise in pressure sensors, by generating synthetic Gaussian noise \(p\).

### B. Bayesian fault isolation

Assume that we have a finite set of possible fault situations, \(f_i, i = 1, \ldots, N\) and a finite set of measuring devices \(x_j, j = 1, \ldots, M\). Assume also that we have a model of the system behavior which allows us the computation of \(M\) residual signals \(r_j, j = 1, \ldots, M\). These time-varying residuals are defined as the mismatch between the measurements and the behavior predicted by the model. When a fault takes place, all the residuals are activated up to some extent.

Given the residuals, the objective is to perform a fault diagnosis procedure in order to identify which fault or faults are behind the observed behavior. In this work, we present a diagnosis procedure based on the Bayesian reasoning. The methodology is explained below.

At every discrete-time instant \(k\), the probability of each fault can be estimated by means of the application of the Bayes rule

\[
P(f_i \mid r(k)) = \frac{P(r(k) \mid f_i) P(f_i)}{P(r(k))}, i = 1, \ldots, N \tag{1}
\]

where \(P(f_i \mid r(k))\) is the posterior probability that fault \(f_i\) had caused the observed residual vector \(r(k) = (r_1(k) \cdots r_N(k))^T\), \(P(r(k) \mid f_i)\) is the likelihood of the residual \(r(k)\) assuming that the active fault is \(f_i\), \(P(f_i)\) is the prior probability for the fault \(f_i\), and \(P(r(k))\) is a normalizing factor given by the total probability law

\[
P(r(k)) = \sum_{i=1}^{N} P(r(k) \mid f_i) P(f_i) \tag{2}
\]

Regarding the prior probabilities, unless we have some additional information, an unprejudiced starting point is to consider all of them equally probable, that is, \(P(f_i) = \frac{1}{N}, i = 1, \ldots, N\). To estimate the likelihood value \(P(r(k) \mid f_i)\) we need to perform a previous calibration task in order to obtain the joint probability density function for each fault in the residual space, \(P(r \mid f_i), i = 1, \ldots, N\). The calibration stage is detailed in the next section. In any case, note that we do not need to assume independency between the residuals.

The application of (1) produces a set of values \(P(f_i \mid r(k)), \sum_{i=1}^{N} P(f_i \mid r(k)) = 1\), that can be used to decide which fault is acting over the system. Note that, at each time sample \(k\), we have information of which is the probability associated to each fault situation. There can be many competing faults, each one with a different probability value. In order to select one of them (the most probable fault), we can take the fault with the highest probability or we can define a threshold value as well.

### C. Recursivity

The results can be improved if (1) is recursively applied, that is, if the posterior \(P(f_i \mid r(k))\) is used as the prior probability for the next time instant. This way, as long as new measurements are available, the probabilities are updated and many of the competing faults can be discarded.

The only drawback is that if any of the faults takes the 1 probability value at any \(k\), then all the remaining faults take the 0 probability value, therefore preventing them to have a future value different from zero due to the recursive application of (1). This drawback can be easily overcome by forcing all probabilities to present a maximum value of, say, 0.99. When a fault \(f_i\) presents the probability \(P(f_i \mid r(k)) > 0.99\), we force it to be \(P(f_i \mid r(k)) = 0.99\) and we can force the remaining faults to be \(P(f_n \mid r(k)) = 1 - 0.99 = 0.01\), \(n = 1, \ldots, N, n \neq i\).

### D. Time horizon

Finally, the result can additionally be improved if a time horizon \(H\) is used. In this case, the posterior probability can be computed on the basis of the \(H-1\) previous computations, that is, to compute \(P(f_i \mid r(k))\), we recursively can apply the following equation

\[
P(f_i \mid r(k-H+n)) = \frac{P(r(k-H+n) \mid f_i) P(f_i \mid r(k-H+n-1))}{P(r(k-H+n))}, \quad i = 1, \ldots, N, n = 1, \ldots, H
\]
where the unprejudiced starting point may be \( P(f_i \mid r(k - H)) = \frac{1}{N}, i = 1, ..., N \).

### III. Calibration

In order to apply the methodology presented in the previous section, we need to perform a first off-line calibration task consisting in obtaining the joint probability density function for each fault in the residual space, \( P(r \mid f_i), i = 1, ..., N \). To do so, we generate a set of residual samples for each fault situation and we obtain the probability density function that best fits to them.

#### A. Residuals independence

In general, better results are obtained if we take into account the correlation between the residual signals. Fig. 3 shows the case where two faults are calibrated by means of Gaussian probability density functions. Fault 1 is better adjusted because it takes into account the cross correlation between residuals \( r_1 \) and \( r_2 \). On the other hand, fault 2 is adjusted by assuming statistic independence between residuals \( r_1 \) and \( r_2 \), therefore the fitting is not so accurate. Note also that other probability distribution families than Gaussian could be used, including multimodal and non-parametric distributions.

#### B. Groups of faults

Sometimes, in practical situations, different faults lead to very similar residual realizations. In such a case, the diagnosis process is more complicated and ends up with a result that indicates that the fault belongs to a certain group of faults whereas the particular fault is not completely isolated (see Fig. 4, fault 1 is introduced in \( k = 25 \) and it cannot be distinguished from faults 2 and 3).

The calibration process allows identifying these situations a priori, for example, by computing the Kullback-Leibler divergence between the obtained probability density functions. If the divergence is below a pre-specified threshold, we can expect that the corresponding faults will be very difficult to isolate.

In these situations, it is interesting to associate a unique PDF for the group of faults. Regarding the prior probability of the group, one can assign a proportional value to the number of faults in the group, i.e., \( P(f_{1,2,3}) = \frac{3}{N} \) or assign the same PDF for all the groups (simple or composed), i.e., \( P(f_{1,2,3}) = \frac{1}{N'} \) with \( N' \) the number of new groups. Working with a joint distribution for the undistinguishable faults allows reducing the number of competing faults and the computational load.

### IV. Case Study

The proposed methodology has been applied to the simplified model of the Hanoi DMA network (depicted in Fig. 5). This model consists of one reservoir, 34 pipes and 31 nodes. Two inner pressure sensors placed in nodes 14 and 30 have been considered (for more details see [19]). The sample time between available measurements is one hour. Single-leak (fault) scenarios in the 31 nodes have been considered. The demand pattern in all demand nodes has been considered known but with some uncertainty as proposed in [20], the leak magnitude has been considered unknown but bounded by an interval (minimum and maximum leak magnitudes) and noise in pressure sensors has been considered too.

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

- A leak magnitude 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 2991.1 \([l/s]\)).
- A study considering noise in pressure measurements with a noise amplitude of \(\pm 5\%\) of the mean value of all pressure residuals.
- A demand uncertainty study considering an uncertainty of \(\pm 5\%\) of the nominal demand node values.
- A study considering the all three uncertainties defined above present in the system at the same time.
A. Calibration

Probability density functions have been calibrated (considering dependence and independence between the different residuals) using a set of data with 200 samples for every leak scenario. Every sample corresponds to two residuals computed with the two sensor measurements and the expected values considering the fault-free scenario. Fig. 6 depicts residuals in the 31 leak scenarios considering demand uncertainty.

![Residuals space with demand uncertainty in Hanoi network.](image)

Fig. 6. Residuals space with demand uncertainty in Hanoi network.

Fig. 7 depicts the PDFs of leak scenario in node 1 when dependence and independence in residuals are considered.

The process of grouping leaks in composed classes described in Section III have been carried out considering the Kullback-Leibler divergence between the 31 obtained PDFs. A threshold of $10^{-3}$ has been chosen. As a result of the grouping process, two composed leak classes that group 4 and 3 single nodes have been obtained. These two composed groups are depicted in Fig. 5. Then, the total of new groups is $N' = 26$.

B. Results

A set of 1000 samples for every leak scenario has been used for validation purposes. The results obtained by the proposed method without recursivity considering dependence and independence between the different residuals in the four different studies have been compared with the ones obtained using the method proposed in [7] where residuals are evaluated by using a kNN classifier. An initial prior probability of $\frac{1}{26}$ for the new 26 groups has been chosen. The results are summarized in Tables I and II. The numbers of Table I indicate the % of well classified leak locations (the highest posterior probability corresponds to the correct hypothesis) defined as “Accuracy”. On the other hand, the numbers of Table II indicate the average position of the correct leak location node among all the different hypothesis in the diagnosis procedure defined as “Rank list”.

As it can be noticed, the proposed Bayesian method considering dependence between residuals provides the best performance in the 4 different studies described before. The improvement of considering dependence (or not) between residuals is significative in case of leak uncertainty. On the other hand, both Bayesian methods can handle more efficiently the dispersion produced by demand uncertainty.

Finally, the effect of the horizon length $H$ in the Accuracy and Rank list of the proposed method considering all the uncertainties together is shown in Figs. 8 and 9. As it can be noticed from these two figures, both indicators improve with the increase of the time horizon length $H$ (i.e. the number of available data).

![Accuracy versus Time horizon length $H$.](image)

Fig. 8. Accuracy versus Time horizon length $H$.

![Pdf of leak scenario in node 1 considering dependence and independence in residuals.](image)

Fig. 7. Pdf of leak scenario in node 1 considering dependence and independence in residuals.
obtained by the proposed method in leak scenario 18. In leak scenario 18 during the first 17 time instants and from $k = 20$ to $k = 22$ the most probable leak provided by the proposed method is leak 3 and the correct leak hypothesis is ranked as the second most probable leak.

V. CONCLUSION

In this paper a leak localization methodology has been presented using pressure residuals and Bayesian reasoning. In a first off-line stage the data obtained from the simulation of the hydraulic model with and without leaks is used to generate the residuals to obtain the PDFs. Also when the PDFs are highly overlapped a grouping method is proposed. The on-line Bayesian reasoning provides the time-dependent posterior probability of every possible leak. The performance of the proposed method has been tested in the Hanoi DMA network in different scenarios and by considering both dependence and independence in the residuals variables. Moreover the performance has been compared with the performance of a previous proposed method that uses a $k$-NN classifier to evaluate the pressure residuals. The effect of the time horizon has also been studied.

VI. ACKNOWLEDGMENT

This work has been partially funded by Spanish Government (Ministerio de Economia y Competitividad) and FEDER 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