Optimal Sensor Placement for Classifier-Based Leak Localization in Drinking Water Networks

Adrià Soldevila¹, Sebastian Tornil-Sin^{1,2}, Rosa M. Fernandez-Canti¹, Joaquim Blesa^{1,2} and Vicenç Puig^{1,2}

Abstract— This paper presents a sensor placement method for classifier-based leak localization in Water Distribution Networks. The proposed approach consists in applying a Genetic Algorithm to decide the sensors to be used by a classifier (based on the k-Nearest Neighbor approach). The sensors are placed in an optimal way maximizing the accuracy of the leak localization. The results are illustrated by means of the application to the Hanoi District Metered Area and they are compared to the ones obtained by the Exhaustive Search Algorithm. A comparison with the results of a previous optimal sensor placement method is provided as well.

Keywords: Water distribution network, sensor placement, fault diagnosis, model-based, classifier.

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 WDN, losses due 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. Then, leak detection and localization are subjects of major concern for water companies. In the case of complex urban WDN it is not easy to deal with these problems. In order to tackle these problems and other problems as pressure control, modern urban WDN are usually divided in District Metered Areas (DMA), where the flow and the pressure at the input are measured [1], [2]. Leakages increase the flow and decrease the pressure measurements at the DMA entrance but leak detection and localization are not trivial problems due to unpredictable variations in consumer demands and measurement noises, as well as long-term trends and seasonal effects. Leak detection can be implemented by means of the analysis of the DMA minimum night flow that can also provide an estimation of the leakage level [1]. However, leak localization usually requires the analysis of more than one measured variable and it is a more complex problem.

In the last years, several techniques have been proposed for leak localization purposes such as transient analysis, parameter estimation techniques, leak sensitivity analysis and artificial intelligence methods. Among them, artificial intelligence methods seem to be a suitable option to deal with the problem of the uncertainty in WDN. In [3] Genetic Algorithms were proposed to solve an optimization problem for simultaneously quantifying and locating water losses. In [4] a method based on the use of Support Vector Machines (SVM) was proposed that analyzes data obtained by a set of pressure control sensors of a pipeline network to locate and calculate the size of a possible leak present in a WDN. More recently, the use of *k*-Nearest Neighbors (*k*-NN) and neuro-fuzzy classifiers for leak localization purposes has been proposed in [5] and [6].

Regarding the kind of sensors, although the use of flow measurements is feasible in large water transport networks, this is not the case in WDN where 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 cheaper and easier to install and maintain.

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. The main problem of optimal pressure sensor placement is that it is a combinatory problem and it is unfeasible to evaluate all the possible sensor locations. In order to deal with this problem the use of Genetic Algorithms (GA) is proposed in [7]–[9] and prior clustering analysis is proposed in [10] and in [11] combined with an efficient branch and bound search.

In this paper, a new sensor placement approach for leak localization in WDN is presented. Given a number of pressure sensors to be installed in the demand nodes of a DMA, the proposed approach provides the locations of the sensors that maximize the accuracy of a leak localization method recently proposed in [5] that combines the use of pressure models with k-NN classifiers. The proposed method requires 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 for all the operating points. Every sensor configuration determines the data that will be able to train the k-NN classifier that will perform the leak localization. In order to tackle the combinatory number of sensor configurations, the use of Genetic Algorithms is proposed to obtain the optimal configuration.

The rest of the paper is organized as follows. Section II

¹Supervision, Safety and Automatic Control Research Center (CS2AC) of the Universitat Politècnica de Catalunya, Campus de Terrassa, Gaia Building, Rambla Sant Nebridi, 22 08222 Terrassa, Barcelona.

²Institut de Robòtica i Informàtica Industrial (CSIC-UPC). Carrer Llorens Artigas, 4-6, 08028 Barcelona.

presents the background of the leak localization combining the use of pressure models and *k*-NN classifiers. In Section III the details of the proposed optimal sensor placement method are provided. Section IV details the application of the method to the Hanoi DMA case study. Finally, Section V draws the main conclusions of the work.

II. BACKGROUND

A. Leak localization pressure residuals and classifiers

In the previous work by the authors [5], it is proposed a on-line leak localization method that relies on the scheme depicted in Fig. 1, based on computing pressure residuals r and analyzing them by a classifier. Residuals are computed as differences between measurements provided by pressure sensors \tilde{p} installed inside the DMA and estimations \hat{p}_0 provided by a hydraulic model simulated under leak-free conditions. The WDN model is built using a hydraulic simulator such as Epanet and it is assumed to be able to represent accurately 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 according to historical consumption records) in the nodes (d_1, \dots, d_N) since in practice real nodal demands (d_1, \dots, 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 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 of these effects into account, the classifier must be able to locate the real leak present in the DMA, 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 instance the position of internal valves and flow and 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.

B. Data generation

The application of the architecture described above (Fig. 1) relies on an off-line work whose main goal is to train and validate a classifier able to distinguish the potential leaks under the described uncertainty conditions. In this process, the data generation stage is critical. 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 (and also validation) data is generated by applying the scheme depicted in Fig. 2, similar to the one presented in Fig. 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. 1. Leak localization scheme.



Fig. 2. Data generation scheme.

The presented scheme is used to:

- Generate data for all possible leak locations, i.e. for all the different nodes in the WDN (\bar{f}_i , i = 1, 2, ..., N).
- For each possible leak location, generate data for different leak magnitudes inside a given range (*f_i* ∈ [*f_i*⁻, *f_i⁺*]).
- Generate sequences of demands $(\bar{d}_1, ..., \bar{d}_N)$ 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 $((\bar{d}_1, ..., \bar{d}_N) \neq (\hat{d}_1, ..., \hat{d}_N))$.
- Take into account the measurement noise in pressure sensors, by generating synthetic Gaussian noise (*ν̄*).

It must be highlighted that the model compute the internal pressures in all the network nodes and that the presented data generation scheme allows generating a complete data set that can be analyzed to determine which pressure measurements are more useful for leak localization purposes.

C. Classifier evaluation

To evaluate the trained classifier the *confusion matrix* Γ can be computed, which summarizes the results obtained when the classifier is applied to a 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 Γ_{ij} indicates how many times a leak in node *i* is recognized as a leak in node *j*. Table I illustrates the concept of the confusion matrix applied to leak localization (in general, to fault isolation).

TABLE I CONFUSION MATRIX Γ

	\hat{f}_1		\hat{f}_i		\hat{f}_N
f_1	$\Gamma_{1,1}$	•••	$\Gamma_{1,i}$	•••	$\Gamma_{1,N}$
÷	÷	÷	÷	÷	÷
f_i	$\Gamma_{i,1}$		$\Gamma_{i,i}$		$\Gamma_{i,N}$
÷	:	÷	:	÷	÷
f_n	$\Gamma_{N,1}$		$\Gamma_{N,i}$	•••	$\Gamma_{n,N}$

In case of a perfect classification, the confusion matrix should be diagonal, with $\Gamma_{ii} = M$, for all $i = 1, \dots, N$ 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 Γ_{ii} indicates the number of times that the leak \hat{f}_i is correctly identified as \hat{f}_i , while $\sum_{j=1}^N \Gamma_{ij} - \Gamma_{ii}$ indicates the number of times that is wrongly classified. The overall accuracy (Ac) of the classifier is defined as:

$$Ac = \frac{\sum_{i=1}^{N} \Gamma_{ii}}{\sum_{i=1}^{N} \sum_{j=1}^{N} \Gamma_{ij}}$$
(1)

III. PROBLEM SOLUTION

A. Problem formulation

The objective of this work is to develop an approach to place a given number of sensors, n, in a DMA of a WDN in order to obtain a sensor configuration with a maximized leak isolability performance when using the leak localization method scheme presented in the previous section. This problem can be recast into the feature selection (also known as variable or attribute selection) problem. The solution of this problem aims at selecting a subset of relevant features (variables) for use in classifier construction. A feature selection algorithm combines a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. However, this is an exhaustive search of the space that is computationally intractable except for small feature sets.

To select a configuration with n sensors, the following binary vector is defined

$$\mathbf{q} = \left[\begin{array}{ccc} q_1 & \cdots & q_N \end{array} \right] \tag{2}$$

where $q_i = 1$ if the pressure in the node *i* is measured, and $q_i = 0$ otherwise (i.e. the vector **q** denotes which sensors are installed).

In order to evaluate the quality of a sensor configuration regarding its capacity to locate a leak at node $i \in \{1, \dots, N\}$, and assuming the case of a single leak, a performance index based on the classified accuracy (1) is proposed.

This performance index depends on the configuration of sensors considered that is parameterized in terms of the binary variable **q** to determine the best selection

$$Ac(\mathbf{q}) = \frac{\sum_{i=1}^{N} \Gamma_{ii}(\mathbf{q})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \Gamma_{ij}(\mathbf{q})}$$
(3)

Note that for a given sensor configuration \mathbf{q} , $100Ac(\mathbf{q})$ is the percentage of correctly located leaks.

Based on the vector \mathbf{q} and the performance index $Ac(\mathbf{q})$ the sensor placement problem can be translated in an optimization problem formulated as follows

$$\max_{\mathbf{q}} Ac(\mathbf{q})$$
s.t.
$$\sum_{i=1}^{N} \mathbf{q}_{i} = n$$
(4)

where $\mathbf{q} \in \{0, 1\}$ is defined in (2) and $n \in \{1, \dots, N\}$ is the number of sensors we want to place.

B. Problem solution

The optimization problem to solve the sensor placement problem, formulated as the classifier feature selection problem described in previous section, is solved using genetic algorithms and implemented using the genetic Algorithm (GA) Toolbox of MATLAB. The GA needs to define a function whose output involves an index to be minimized (or a negative index if the index is wanted to be maximized at the current case).

The training matrix $\mathbf{M_T}$ has N + l + 1 columns where the first N columns are the node measurements, the next l columns are the l added attributes (e.g., the total water inflow) used by the classifier, and the last column corresponds to the label where each data scenario belongs; the number of rows corresponds to the total number of data scenarios used to train the classifier. The validation matrix $\mathbf{M_V}$ has the same column format (N + l + 1), and the number of rows corresponds to the total number of data scenarios used for validation purposes.

The pseudo-code of the algorithm is shown in Algorithm 1. First, we initialize the variables of the GA (line 1) including the bit string type population, the tolerance as 10^{-6} , the population size p and the elite count as 0.05p

Algorithm 1 Sensor placement based on Genetic Algorithms

- **Require:** A training matrix M_{T} , a validation matrix M_{V} . The number of sensors n, the number of nodes N and the population size p.
- Ensure: A near-optimal sensors configuration q with error index Ac_{max} .
 - 1: $init \leftarrow InitVarGA()$
 - 2: $constraint \leftarrow SetConstraints()$
 - 3: Inputs: *init*, constraint, M_T , M_V , p, N.
 - 4: while An optimization criterion is not reached do
 - GA based search: 5:
- Generate I matrix of size $(p \times N)$ where each row 6: is a member of a generation.
- for $k = 1, \cdots, p$ do 7:

 $\mathbf{q} \leftarrow I(p)$ 8:

- $C(\mathbf{q}) \leftarrow Train(\mathbf{M}_{\mathbf{T}}(\mathbf{q}))$ 9:
- $\Gamma(\mathbf{q}) \leftarrow Validate(C(\mathbf{q}), \mathbf{M}_{\mathbf{V}}(\mathbf{q}))$ $Ac(\mathbf{q}) \leftarrow \frac{\sum_{i=1}^{N} \Gamma_{ii}(\mathbf{q})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \Gamma_{ij}(\mathbf{q})}$ 10:
- 11:
- end for 12:
- Find $\{\mathbf{q}^k, Ac^k\}$ such that $Ac^k = \max_{\mathbf{q}}(Ac(\mathbf{q})).$ 13:

14: end while

(at least one member survives) in order to save part of the previous analyzed results. Then, we declare the search constraints (line 2) being n the constraint of the set of possible solutions for each variable and the number of sensors. Then, in the optimization process (lines 4 to 14), an initial matrix with random sensor positions is delivered by the GA (line 6) which is tested according to the objective function (line 11). The sensor placement is based on the construction of binary vectors q where the presence of a "one" represents a sensor located in the correspondent node. This vector **q** allow to select the adequate columns of the matrices M_T and M_V in order to train (line 9), validate (line 10) and compute Ac (line 11) according to the selected nodes to be measured. Once we have the Ac value for all members of the matrix I, we look for the maximum value (line 13) and if the maximum value accomplish the optimization criteria (in this particular case, if the variation between generations after 50 generations is less than the tolerance value), then the optimization is finished given as result the sensor placement used to obtain the last Ac result.

IV. CASE STUDY AND RESULTS

The proposed method is tested in the Hanoi DMA network placed in Vietnam. This network consists in 31 nodes, 36 pipes and one reservoir, as depicted in Fig. 3. Previous works [8] showed that the best number of inner pressure sensors to place in the network (taking into account the cost of install and maintain the sensors) is two. The daily pattern demand is simulated as the one depicted in Fig. 4, and the demand distribution inside the network is considered known but with some uncertainties as proposed



Fig. 3. Hanoi topological network.



Fig. 4. Example of a daily flow consumption.

in [12]. Only the single fault case (one leak present at each time) is considered.

The two pressure sensors, and the total amount of water entering into the network are measured every 10 minutes. Instead of using the six measurements available each hour, their average is used in order to avoid outliers and reduce the uncertainty in the data sets.

The uncertainties considered (using bounded uniformly distributed random numbers) during the sensor placement in this network are the following:

- A leak magnitude uncertainty in the range between 25 and 75 [l/s] is used.
- The noise in the measurements is considered of the order of \pm 5 % of the average value of all pressure residuals.
- The uncertainty in the nodal demands is considered of \pm 10 % of the nominal value.

The k-NN classifier uses as attributes the two pressure residuals and the total water inflow, the k = 1 is chosen without any particular criteria and the distance rule used is the Euclidean distance.

This network is small enough to apply the Exhaustive Search (ES) method, i.e. to compute and evaluate all the possible combinations, since the use of two sensors lead the number combinations (eliminating repeated sensors in a placement) up to 465. The ES results are used to compare against the results obtained using the GA method proposed.

Every data set is divided in three parts (training, valida-

Method			
		MAS [NTA]	14, 28 [4]
ES	A	verage Time [s]	1874.6
	Aver	rage Accuracy [%]	41.24
	PS = 5	MAS [NTA]	18, 24 [3]
GA		Average Time [s]	195.1
		Average Accuracy [%]	41.07
	PS = 10	MAS [NTA]	12, 28 / 14, 28 [4]
		Average Time [s]	351.33
		Average Accuracy [%]	41.22
	PS = 20	MAS [NTA]	14, 28 [4]
		Average Time [s]	504.43
		Average Accuracy [%]	41.24

TABLE II Results using the ES and the GA methods.

tion and testing data sets). The training data set has four samples per each leak and hour (96 samples for each class, corresponding to four days of data). This set is used to train the k-NN classifier. The validation data set consists in two samples per each leak and hour (48 samples for each class, corresponding to two days of data) and it is used to calculate the "Accuracy" value used in the objective function in the GA method and the ES method to find the best suitable sensor placement. Finally, the testing data set has the same size as the validation data set and it is used to obtain the "Accuracy" value to compare the results obtained using the different methods.

For the GA method different population sizes are used to evaluate the effect of this parameter in the optimization. It must be taken into account that the evaluation of the objective function for each possible combination is performed only once and it is stored for the case that the combination appears again in order to improve efficiency.

Ten different data sets (including training, validation and testing data sets) are used with the characteristics described before. The results are summarized in Table II, where the terms "PS" refers to the population size, "MAS" to the combination of sensors that most appear, and "NTA" the number of times that the combination appears. The results have been all obtained in a PC with an INTEL(R) CORE(TM) i7-4720HQ CPU @ 2.60 [GHz], 8 [GB] of memory RAM and a Windows 10 Home 64 bits operative system and using the MATLAB 2015a software.

From these results, it can be seen that the GA method performs differently depending on the population size parameter, but the "Accuracy" remains high in all cases while the computational time grows fast as the population size increases. In the case that the population size is 20 members, the GA method finds 10 of 10 times (for each data set considered) the same sensor placement as the ES method, so it seems to have a closest performance to the ES method and achieve the optimum or near optimum result.

A previous optimal sensor placement [8] based on the Angle method [13] is used to compare how the proposed method improves. The criteria used for this placement is not optimum for the k-NN classifier method but still it is a good criteria, since the aim of this method is to



Fig. 5. Comparison of the accuracy results in Hanoi DMA network.



Fig. 6. Hanoi residuals space (each color represents a different leak) for the sensor placement obtained using the Angle method.

maximize the angle (cosine) value between residuals of different classes, which objective is to separate the classes in the residual space to obtain a better accuracy (same final objective). The application of this method results in a sensor placement with nodes 12 and 21. Note that the uncertainties and daily pattern consumption are different to the ones presented and used in this paper.

Fig. 5 shows a comparison of the performance (accuracy indicator) of the method of sensor placement presented (k-NN GA method, sensors 14 and 28) and the compared method (angle method, sensors 12 and 21) using a data set whose optimum sensor placement is the same as the obtained with the k-NN GA method, and for a time horizon H (number of hourly samples available to perform the leak localization).

An example of the residuals space in one of the training data (data set number six, which is the same set of data that has been used to obtain the results depicted in Fig. 5) is depicted in Fig. 6 for the sensor placement obtained using the Angle method and Fig. 7 for the one obtained with the GA method, in this case, with the three population sizes.

V. CONCLUSIONS

In this paper an optimal sensor placement method for placing a given number n of pressure sensors in WDNs has been presented. The obtained sensor configuration is optimal in the sense that it maximizes the leak isolability when



Fig. 7. Hanoi residuals space (each color represents a different leak) for the sensor placement obtained using the *k*-NN GA method presented.

using a k-NN leak localization method. In order to tackle the combinatory problem of optimal sensor placement, the use of Genetic Algorithms has been proposed. The performance of the proposed method has been illustrated by means of the application to the Hanoi District Metered Area. For a moderate population size, the simulation results show that the Genetic Algorithm provides the same performance than the Exhaustive Search Algorithm whereas the computation load decreases significantly. Thus this method becomes suitable for networks of growing complexity. Additionally, compared to the sensor placement obtained by the Angle method, the use of classifiers allows the direct introduction of different sources of uncertainty and leads to improved isolability results. Future work includes the optimization of the number of sensors and the application of the methodology to a real WDN.

VI. ACKNOWLEDGMENT

This work has been partially funded by the Spanish Government (MINECO) through the project CICYT ECOCIS (ref. DPI2013-48243-C2-1-R), by MINECO and FEDER through the project CICYT HARCRICS (ref. DPI2014-58104-R), by MINECO through the grant IJCI-2014-2081, by EFFINET grant FP7-ICT-2012-318556 of the European Commission, by the DGR of Generalitat de Catalunya (SAC group Ref. 2014/SGR/374) and by the AGAUR of Generalitat de Catalunya through the grants FI-DGR 2015 (ref. 2015 FI_B 00591) and 2014PDJ00102.

REFERENCES

- R. Puust, Z. Kapelan, D. A. Savic, and T. Koppel, "A review of methods for leakage management in pipe networks," *Urban Water Journal*, vol. 7, no. 1, pp. 25–45, 2010.
- [2] M. F. Lambert, A. R. Simpson, J. P. Vítkovský, X. J. Wang, and P. J. Lee, "A review of leading-edge leak detection techniques for water distribution systems," in 20th AWA Convention, Perth, Australia, 2003.
- [3] Z. Y. Wu and P. Sage, "Water loss detection via genetic algorithm optimization-based model calibration," in *Systems Analysis Symposium.* ASCE, 2006, pp. 1–11.
- [4] J. Mashford, D. De Silva, D. Marney, and 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.

- [5] L. Ferrandez-Gamot, P. Busson, J. Blesa, S. Tornil-Sin, V. Puig, E. Duviella, and A. Soldevila, "Leak localization in water distribution networks using pressure residuals and classifiers," *IFAC-PapersOnLine*, vol. 48, no. 21, pp. 220 – 225, 2015.
- [6] D. Wachla, P. Przystalka, and W. Moczulski, "A method of leakage location in water distribution networks using artificial neuro-fuzzy system," *IFAC-PapersOnLine*, vol. 48, no. 21, pp. 1216 – 1223, 2015.
- [7] R. Pérez, V. Puig, J. Pascual, J. Quevedo, E. Landeros, and A. Peralta, "Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks," *Control Engineering Practice*, vol. 19, no. 10, pp. 1157–1167, 2011.
- [8] M. V. Casillas, V. Puig, L. E. Garza-Castañón, and A. Rosich, "Optimal Sensor Placement for Leak Location in Water Distribution Networks Using Genetic Algorithms," *Sensors*, vol. 13, no. 11, pp. 14 984–15 005, 2013.
- [9] M. A. Cugueró-Escofet, V. Puig, J. Quevedo, and J. Blesa, "Optimal pressure sensor placement for leak localisation using a relaxed isolation index: Application to the barcelona water network," *IFAC-PapersOnLine*, vol. 48, no. 21, pp. 1108 – 1113, 2015.
- [10] J. Blesa, F. Nejjari, and R. Sarrate, "Robust sensor placement for leak location: analysis and design," *Journal of Hydroinformatics*, vol. 18, no. 1, pp. 136–148, 2016.
- [11] R. Sarrate, J. Blesa, F. Nejjari, and J. Quevedo, "Sensor placement for leak detection and location in water distribution networks," *Water Science and Technology: Water Supply*, vol. 14, no. 5, pp. 795–803, 2014.
- [12] P. Cugueró-Escofet, J. Blesa, R. Pérez, M. Cugueró-Escofet, and G. Sanz, "Assessment of a leak localization algorithm in water networks under demand uncertainty," *IFAC-PapersOnLine*, vol. 48, no. 21, pp. 226 – 231, 2015.
- [13] M. V. Časillas, L. E. Garza-Castañón, and V. Puig, "Extendedhorizon analysis of pressure sensitivities for leak detection in water distribution networks: Application to the Barcelona network," ... *Conference (ECC)*, 2013..., no. 1, 2013.