Topological analysis of water distribution networks for optimal leak localization

Débora Alves[1,3], **Joaquim Blesa**[1.2.4], **Eric Duviella** [3] **and Lala Rajaoarisoa**[3]

[1] Supervision, Safety and Automatic Control Research Center (CS2AC) of the Universitat Politècnica de Catalunya, Campus de Terrassa, Gaia Building, 08222 Terrassa, Barcelona, Spain
[2] Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Carrer Llorens Artigas, 4-6, 08028 Barcelona, Spain
[3] IMT Nord Europe, Univ. Lille, CERI Digital Systems, F-59000Lille, France
[4] Serra Húnter Fellow, Universitat Politècnica de Catalunya (UPC),Automatic Control Department (ESAII), Eduard Maristany, 16 08019, Barcelona, Spain

adeboracris@gmail.com

Abstract.: This paper introduces two methodologies to provide an optimum sensor deployment layout, one based on a model-based approach and the other entirely data-driven. The first method is formulated as an integer optimization problem, an optimization criterion consisting of minimizing the average topological distance. The second method is a new methodology to provide an optimum sensor placement regarding how many sensors to install without using hydraulic information but just exploiting the knowledge of the topology of the Water Distribution Networks. The method uses the Girvan-Newman clustering algorithm to ensure complete coverage of the network and the study of the installation of pressure sensors in the central nodes of each group, selected according to different metrics of topological centrality. The approach is illustrated in the Modena WDN.

Keywords. Data-Driven, Leak Localization, Sensor Placement, Water Distribution Network

1. Introduction

Water Distribution Networks (WDN) are complex networks due to their size (thousands of pipes) and hydraulic behavior due to their nonlinearity. One of the main problems in these networks is leaks that may appear in the system due to different factors, like as weak joints, water hammers, utility construction or excavation, seasonal temperature changes, and other things. Therefore, a significant effort has been made to reduce the impact of leakage on the network, such as the leak localization study that indicates which area in the WDN there may be a leak.

Several works on leak localization were released by applying model-based approaches and commonly used demand-driven (DD) hydraulic simulators. For example, in [1], the research is based on the analysis of pressure residues. Moreover, in [2], the authors use hydraulic models with AI methods. The results based on hydraulic models are excellent. Nevertheless, the main difficulties characterize are the calibration of accurate models and data availability for all possible complex scenarios.

Thinking about these challenges, recent studies have proposed analyzing the problem using data-driven methods [3,4] that combine the use of standard operation data and topological information. The particular method in [5] studies the effect of the extra flow when a leak occurs in the pressure sensors presented in the network. It aims at developing a relative incidence of a leak using network topology correlated with the flow and pressure measurement.

An element that has the potential to significantly improve the localization of leaks is the sensor pressure configuration. In the last years, several strategies that tackle the problem of optimal sensor placement in WDNs for leak localization have been proposed. As a branch and bound searches [6], Genetic Algorithms [7], feature selection techniques [8], and game theory approaches [9].

This work presents two new methodologies for sensor placement in the WDN, one using hydraulic simulation formulated as an integer optimization problem solved with a Genetic Algorithm (GA). And the other uses only the topological network information to improve the leak localization methods that use residual analysis [1,3,5]. The second approach aspires to simplify the problem of sensor placement by eliminating the need to calibrate the hydraulic water models and reducing the computational burden. The methodology is based on the complex network theory applying the graph approach with the hydraulic information to represent the WDN.

The rest of the document is organized as follows: Section 2 presents the leak localization methodology. Section 3 presents the sensor placement algorithms proposed in this work. Section 4 shows the application and the results obtained in a real water distribution network. Finally, Section 5 concludes this work.

2. Leak localization

Using pressure measurements, the leak location methodology aims to detect and isolate leaks in a water distribution network. Normally, the methods are triggered when a leak is detected. Leak detection is usually done using inlet flow analysis. Considering that inlet pressure and flow sensors and other pressure sensors in inner nodes are installed in the WDN. Leak localization can be carried out by employing the analysis of pressure residuals generated by the comparison of inner pressure measurements and leak-free pressure estimations as

$$r_i = \hat{p}_i(c) - p_i(c) \qquad i = 1, \ldots, s \tag{1}$$

where r_i, $\hat{p}_i(c)$ and $p_i(c)$ are the residual, leak-free pressure estimation, and pressure measurement at inner node *i*. *c* is the operating condition defined by inlet measurements and *s* is the number of inner sensors installed in the WDN. Leak-free pressure estimations $\hat{p}_i(c)$ can be computed by physical models or through of historical data. If a physical model of the WDN is available, a leak sensitivity matrix Ω

$$\Omega = \begin{pmatrix} \frac{\partial r_1}{\partial f_1} & \cdots & \frac{\partial r_1}{\partial f_n} \\ \cdot & \cdots & \cdot \\ \frac{\partial r_s}{\partial f_1} & \cdots & \frac{\partial r_s}{\partial f_n} \end{pmatrix} \tag{2}$$

where $\frac{\partial r_i}{\partial f_j} = \hat{p}_i(c) - \hat{p}_i^j(c) \quad i = 1, \ldots, s \quad j = 1, \ldots, n$ with $\hat{p}_i^j(c)$ is the pressure in node *i* considering a leak in node j denoted as f_j. Then, leak localization can be formulated as the maximum correlation between the observed residuals and the different leak hypothesis

$$\underset{j\in\{1,\ldots,n\}}{\arg\max}\ \frac{\omega_{\bullet j}\cdot\boldsymbol{r}}{\left\|\omega_{\bullet j}\right\|\left\|\boldsymbol{r}\right\|} \tag{3}$$

where $\omega_{\bullet j}$ is the j^{th} column of a sensitivity matrix (2) and $\boldsymbol{r}$ is the residual whose components are computed in equation (1). Alternatively, if it is not available any hydraulic model of the WDN the leak localization method can be formulated as the maximum residual component

$$\underset{i\in\{1,\ldots,s\}}{\arg\max}\ \{r_i\} \tag{4}$$

The main disadvantage in the use of leak localization in equation (4) compared with leak localization in equation (3) is that the result of the leak localization is not a node, but a cluster related to one of the s inner pressure sensors. But as it is very simple, and it does not require any physical model, it is a good reference point to develop new data-driven leak localization methods. As in [1,5], where topological information was used to formulate equation (4) at the node level. In this work equation (4) is used to be the leak localization method.

3. Sensor placement

This work aims to develop an approach to placing a given number of sensors, s, in a WDN to obtain a sensor configuration with a maximized leak localization performance. In order to cope with the combinatory complexity of the sensor placement problem, following the ideas of [10], a two-step suboptimal search algorithm is proposed:

STEP 1: Divide the nodes of the WDN into s clusters $C=\{C_1,\ldots,C_s\}$, i.e., a cluster for every sensor to be installed in the WDN.

STEP 2: Choose a node among all nodes of a cluster as the optimal place to install a sensor.

In order to carry out the STEP 1, the WDN can be represented as a directed graph $G=(V,E)$, with V as the set of vertices that represents the n connections between the components of the network (junctions, reservoirs, and tanks), and E are the edges, which represent the m links (pipes, valves, and pumps) in the network. The edges are associated with a cost value based on the friction loss in pipes of the Hazen-Williams formula, that is, the pipe length divided by the pipe diameter, to guarantee a model closer to the real behavior of the water system. The Girvan-Newman (GN) clustering method [11] is proposed for STEP 1, GN clustering is an algorithm that focuses on edges mostly between communities, so clusters are defined by progressively removing edges from the original graph according to edge betweenness, which measures the importance of an edge in a network by counting the number of shortest paths that run through it.

For STEP 2, two methods were developed: the first being the model-based approach which uses hydraulic models formulated as an integer optimization problem. This approach is only possible to use if the hydraulic model has high credibility. Furthermore, the second is a data-driven approach to locate sensors at the topologically most essential nodes of each cluster, ensuring a spatially uniform distribution of sensors.

3.1. Model-based approach

STEP 1 gives the information of the $C=\{C_1,\ldots,C_s\}$ clustering with the numbers of nodes that make up each one

$$\begin{aligned} X_1 &= \left\{1,\dots,C_1^n\right\} \\ X_2 &= \left\{C_1^n+1,\dots,C_1^n+C_2^n\right\} \\ &\vdots \\ X_s &= \left\{\sum_{i=1}^{s-1} C_i^n+1,\dots,\sum_{i=1}^{s-1} C_i^n+C_s^n\right\} \end{aligned} \tag{5}$$

where $X_1,\dots,X_s$ are the groups of nodes that contain the cluster C, they are organized in crescent order with C^n represent the last node of each clustering and C_s^n value is equivalent to the total n node numbers of WDN. The goal in STEP 2 is to choose one node in each X to be the place to install a sensor. A performance index, Average Topological Distance (ATD) that displays the information on the node's distance between the node predicted as leaking and the actual node with the leak can be minimized to perform the optimal sensor placement using the equation (3).

To calculate the ATD is first necessary to create a matrix containing the minimum topological distance (in nodes) $D \in \mathbb{R}^{nxn}$. And the confusion matrix Γ depicted in Table 1 is used to assess the performance of equation (3). The rows of this matrix correspond to the leak scenario and the columns to which the leak is located by the leak localization method.

Table 1. Confusion matrix Γ.

	$\hat{l}_1$	$\dots$	$\hat{l}_i$	$\dots$	$\hat{l}_n$
l_1	$\Gamma_{1,1}$	$\dots$	$\Gamma_{1,i}$	$\dots$	$\Gamma_{1,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
l_i	$\Gamma_{i,1}$	$\dots$	$\Gamma_{i,i}$	$\dots$	$\Gamma_{i,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
l_n	$\Gamma_{n,1}$	$\dots$	$\Gamma_{n,i}$	$\dots$	$\Gamma_{n,n}$

In this way, the ATD can be calculated as:

$$f(x) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\Gamma_{i,j}D_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\Gamma_{i,j}} \tag{6}$$

where $x = \left\{x_1,\dots,x_s\right\}$ is the set of s sensors, with the constraint of $x_1 \in X_1,\dots,x_s \in X_s$. In addition, the optimal value is the minimal distance. Based on the performance index f the sensor placement problem is cast as an integer optimization problem formulated as:

$$\begin{aligned} \min_x : &\qquad f(x) \\ s.t: &\qquad x_i^{\,l} \le x_i < x_i^{\,u}, \quad i = 1,\cdots,s \\ &\qquad x \in \mathbb{Z}^n \end{aligned} \tag{7}$$

where x_i^l, x_i^u is the lower and upper bounds based on the equation (5). It should be noticed that the solution of the previous optimization algorithm provides the best sensor location when the operating conditions are similar to the one used to evaluate residuals in equation (1). If the operating conditions are different, the optimal sensor location could vary.

3.2. *Data-driven approach*

As explained in the previous section, STEP 2 aims to select nodes in each $X_1,...,X_s$ defined in equation (5). The core idea of the present section is to locate sensors without using any hydraulic simulations because data availability is often limited or not suitable. In addition, it reduces the computational burden.

Thus, in order to define a criterion to approach the sensor placement problem, also in the case of unavailable or incomplete hydraulic information on the network, the topology most central nodes of each cluster are considered suitable sensors locations. For this purpose, three indicators of the importance of the nodes were used to select the positioning of the sensors:

- Closeness centrality uses the inverse sum of the distance from a node to all other nodes in the graph, the more central a node is, the closer it is to all other nodes.

$$c_c(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)} \tag{8}$$

  where $d(i,j)$ is the distance between vertices i and j. And n is the number of nodes in the graph/clustering;
- Betweenness centrality: measures how often each graph node appears on a shortest path between two nodes in the graph

$$c_{st}(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{9}$$

  where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(i)$ is the number of those paths that pass through i;
- Eigenvector centrality uses the eigenvector corresponding to the largest eigenvalue of the graph adjacency matrix.

$$A_s x = \lambda_{\max} x \tag{10}$$

  where A_s is the adjacency matrix of the subgraph G and $\lambda_{\max}$ is the largest eigenvalue. It computes the centrality for a node based on the centrality of its neighbors, according to the coordinates $e_c(v)$ of the eigenvector $x = e_c$, associated with the largest eigenvalue of $\lambda_{\max}$ matrix A_s.

3.3. *Performance indicators*

The proposed leak localization performance indicators to assess the sensor placement optimality in this work are the following:

- Average topological distance (ATD): represents the distance in nodes between the centroid predicted as leaking with the true node leaks. The ATD index that presents a minimum value is preferable. First, it is necessary to create a matrix containing the minimum topological distance (in nodes), i.e., equation (6). This index is a suitable parameter for analyzing the improvement of the leak localization method;

- F1-score is the weighted average of precision and recall, where precision is the analysis of all positive predictions, how many are positive, and recall is the study of real positive cases, how many are predicted positives. The F1 score is a good indicator of imbalanced data [11];
- Cohen's Kappa: represents the degree of accuracy and reliability, which is the difference between the observed overall accuracy of the model and the overall accuracy obtained by chance. It is a more practical measure to use on problems with an imbalance in the classes. The kappa has a range from −1 to +1. Values between 0.6 and 0.8 are considered good [12].

The indexes proposed to analyze the performance of sensor placement were presented to analyze the improvement of sensor location using the average topological distance of the coverage area where the sensor can identify a leak. Moreover, the ATD index is an excellent index to measure the improvement of leak localization methods. Furthermore, the F1 score and the Cohen's Kappa were suggested to analyze the GM clustering method and the node importance centroids according to the residues, being the higher these indexes are, the better the GM performance.

4. Case study

The case study selected to test the performance is the reduced model of the real water distribution network of the Italian city Modena. This large-scale network comprises 268 junctions (nodes) connected through 317 pipes and served by four reservoirs.

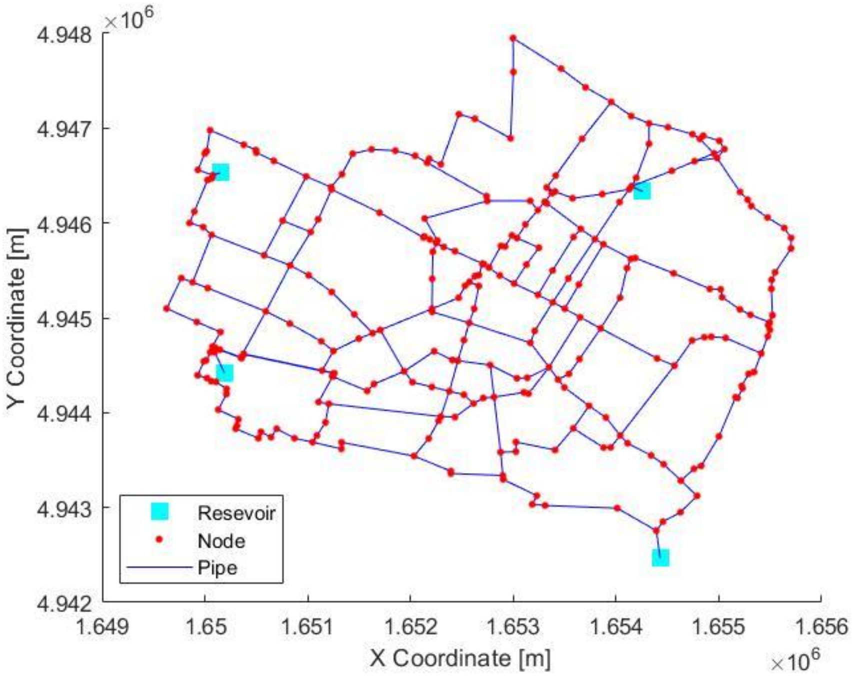

Figure 1. Simplified Modena topological WDN.

EPANET hydraulic simulator was used to generate a leak lasting 72h scenario data to analyze the performance of the proposed method. The following simulation conditions were used:

- to reduce the uncertainty in the data, samples were collected every 10 minutes and filtered to 24 hours values;
- the uncertainty of 10% (normal distribution) of the nominal demand value was considered;
- the leak size was randomly selected, with 3 to 6 l/s representing 1 to 2.5% of the network consumption.

The minimization of the optimization presented in section 3.1 is carried out using Genetic algorithms (GA) based on principles of natural genetics and natural selection [13,14]. The GA can be used in the context of sensor placement in WDN to find the near-optimal placement of these sensors for leak localization. In that case, a chromosome corresponds to the possible presence or absence of a sensor at a given node.

Table 2 shows the results obtained in three different scenarios: with 3, 5, and 10 possibilities of sensor placement. The number of nodes chosen to have a sensor is exhibited in all scenarios for the GA solution and each node importance method. As the optimization of equation (7) is based on the optimization of the ATD index, the results presented in Table 2 show that the solution obtained by the GA will be the best, in any case, even if the proposed data-driven methodology is not the optimal solution of the sensor network, the obtained values are not far from those of GA. Note that even getting the best optimal value of the ATD, the solution obtained by the GA does not guarantee the best result for the F1 score and the kappa. The Eigenvector centrality is the results that present the worst results, principally in the scenario with three sensors, having the Cohen's Kappa with the worst value, inferior a 0.6.

On the other hand, the Betweenness and the Closeness centrality had a similar result. However, analyzing the ATD index, the Betweenness metric improves the scenario with 3 and 5 sensors. In the scenario with ten sensors, the Closeness centrality has the better performance. Whereas in a general case, the Betweenness centrality is the best choice in the data-driven methodologies.

Table 2. Average evaluation metrics.

Criterion	Nodes with sensors	ATD	F1 score (%)	Kappa
3 sensors scenario				
GA	**88 109 207**	**6.82**	**82.43**	**0.74**
Closeness	91 147 207	7.50	82.01	0.71
Betweenness	7 63 109	7.37	79.63	0.70
Eigenvector	4 83 119	8.05	67.48	0.48
5 sensors scenario				
GA	**5 41 80 110 164**	**5.80**	**47.01**	**0.71**
Closeness	49 91 135 147 207	6.50	82.01	0.75
Betweenness	7 63 49 91 135	6.54	81.91	0.75
Eigenvector	4 83 92 119 134	6.70	75.64	0.65
10 sensors scenario				
GA	**10 31 47 78 129 171 129 183 225 258**	**4.45**	**28.97**	**0.62**
Closeness	11 31 49 83 91 105 137 129 180 258	4.70	73.36	0.68
Betweenness	1 11 31 35 49 83 91 129 135 180	4.68	69.60	0.64
Eigenvector	1 4 31 83 92 102 119 129 134 180	4.71	83.36	0.79

Figure 2 shows the result of the second scenario with 5 sensors. Each cluster C is highlighted with a different shade. The position of the sensors obtained with the proposed methods is shown with a circle with different colors. Clustering generated with GN provides a division focused on edges between communities that do not guarantee a homogeneous distribution of nodes. As the nodes are more communicated, the effect on the residue in a leak between these nodes will affect more the sensor installed in that region.

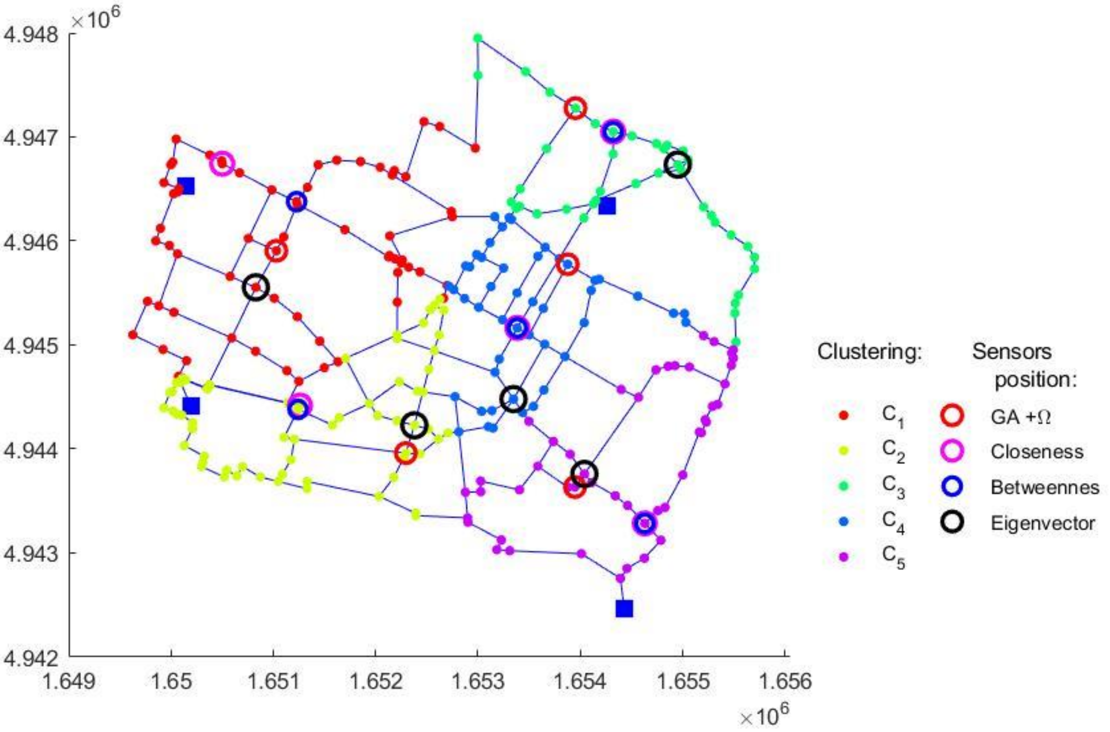

Figure 2. WDN of Modena and the four sensor layouts according to the three topological centrality metrics and GA solution.

5. Conclusions

The quality of sensor placement in WDSs impacts leak identification efficiency, and sensor-placement optimization remains one of the top issues in related research. In this work, a new full sensor placement method utilizing just the topological information of the WDN based on the high connection density of the graph representation of the system in association with the node importance has been presented in this study. It employs only the topological characteristics of a WDN, combining the identification of clusters of nodes and topological centrality metrics for the design without carrying out any hydraulic simulation. It was proposed to provide a tool specially adapted for the frequent case where only partial information about the system is available.

In addition, a new approach to sensor placement that minimizes the average topology distance of leak isolability has been proposed, formulated as an integer optimization problem. However, the method uses a hydraulic model to simulate all node leaks based on the system's demand pattern and uncertainties.

The proposed approaches have been explained, and an example is presented using the Modena Network simplified version of the real WDN as a case study. They demonstrate that the methodology using only the system topology information obtained a good result, ideal for cases where partial details on the system are available. To define the most appropriate procedure for the design of sensor placement, future work will investigate how different objective functions can improve the selection, and the effect of other stressing conditions (i.e., sensor failures) in the network can change the result of sensor placement.

Acknowledgments

This work has been partially funded by L-BEST project (Ref. PID2020- 115905RB-C21) from the Spanish Ministry of Science.

References

[1] Soldevila, A., Fernandez-Canti, R.M., Blesa, J., Tornil Sin, S., and Puig, V. (2016). Leak localization in water distribution networks using model-based bayesian reasoning. In 2016 European Control Conference (ECC),1758–1763. IEEE.

[2] Javadiha, M., Blesa, J., Soldevila, A., and Puig, V. (2019). Leak localization in water distribution networks usingdeep learning. In 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), 1426–1431. IEEE.

[3] Soldevila, A., Boracchi, G., Roveri, M. et al. Leak detection and localization in water distribution networks by combining expert knowledge and data-driven models. Neural Comput & Applic (2021)

[4] Sun, Congcong, et al. "Leak localization in water distribution networks using pressure and data-driven classifier approach." Water 12.1 (2020): 54

[5] Alves, D.; Blesa, J.; Duviella, E.; Rajaoarisoa, L. Robust Data-Driven Leak Localization in Water Distribution Networks Using Pressure Measurements and Topological Information. Sensors 2021, 21, 7551.

[6] 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.

[7] D. B. Steffelbauer, D. Fuchs-Hanusch, Efficient sensor placement for leak localization considering uncertainties, Water Resources Management 30 (14) (2016) 5517–5533.

[8] Soldevila, A.; Blesa, J.; Tornil-Sin, S.; Fernandez-Canti, R.M.; Puig, V. Sensor placement for classifier-based leak localization in water distribution networks using hybrid feature selection. Computers & Chemical Engineering 2018,108, 152 – 162.

[9] G. Arbesser-Rastburg, D. Fuchs-Hanusch, Serious sensor placement—optimal sensor placement as a serious game, Water 12 (1) (2020).

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

[11] Richard, L.J.; Koch, G.G. An Application of Hierarchical Kappa-Type Statistics in the Assessment of Majority Agreement Among Multiple Observers. Biom. JSTOR 1977, 1977, 363–374

[12] C. J. Van Rijsbergen. Information Retrieval. Butterworth-Heinemann, 1979. Romano, M.; Woodward, K.; Kapelan, Z. Statistical Process Control Based System for Approximate Location of Pipe Bursts and Leaks in Water Distribution Systems. Procedia Engineering 2017, 186, 236–243.

[13] J. R. Koza, "Survey of genetic algorithms and genetic programming," in In In Proceedings of the Wescon 95 - Conference Record: Micro-electronics, Communications Technology, Producing Quality Products, Mobile and Portable Power, Emerging Technologies. IEEE Press,1995, pp. 589–594

[14] Girvan, M., and M. E. Newman. 2002. "Community structure in social and biological networks." Proc. Natl. Acad. Sci. 99 (12): 7821–7826. https://doi.org/10.1073/pnas.122653799.