A comparison between model-based and data-driven leak localization methods ⋆

Luis Romero-Ben * **Joaquim Blesa** *,**,*** **Gabriela Cembrano** * **Vicenç Puig** *,**

* *Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain (e-mail: luis.romero.ben@upc.edu, gabriela.cembrano@upc.edu).*
** *Supervision, Safety and Automatic Control Research Center (CS2AC) of the Universitat Politècnica de Catalunya, Rambla Sant Nebridi 22, 08222 Terrassa, Spain (e-mail: joaquim.blesa@upc.edu, vicenc.puig@upc.edu)*
*** *Serra Húnter Fellow, UPC, Automatic Control Department (ESAII), Eduard Maristany 16, 08019 Barcelona, Spain*

Abstract: Water networks are crucial infrastructures for the sustainability of modern cities, and hence their proper operation is of great importance. This includes the fast detection, localization and repair of leaks, which may produce major water losses. This paper presents a comparison between two leak localization methods, which belong to opposite categories: model-based and data-driven. To this end, their main characteristics are reviewed, highlighting their differences and advantages/drawbacks, to finally display several results whose discussion allows to draw important conclusions for the future of the research field.

Keywords: Water distribution, fault localization, model-based, data-driven, comparison.

1. INTRODUCTION

The effective management of leakage in water distribution systems (WDS) is of upmost importance for water utilities. Liemberger and Wyatt (2019) estimated water losses of 126 billion cubic meters of water per year worldwide. This illustrates the high costs of water leaks, justifying the interest in leak detection and/or localization methodologies.

This research field has been widely studied during the years, starting from hardware-based methods that use devices to solve the task, yielding accurate results although their usage is limited to small areas. Later, software-based approaches appeared, using data from sensors, models and algorithms to detect/locate leaks. This group is normally divided into two categories: model-based and data-driven.

Model-based methods require a hydraulic model of the WDS, calibrated in terms of both physical properties and demands, to perform simulations and compare the computed data with actual measurements. An early attempt of this kind was proposed in Pudar and Liggett (1992), which studies the effect of each possible leak on every node of the network. A well-known localization methodology was proposed in Perez et al. (2014), using a fault signature matrix to discern the most probable location of the leak represented by the measured pressures. Recently, Steffelbauer et al. (2022) proposed a complete calibration, detection and localization methodology, with the latter based on a residual projection strategy with model update, in order to operate over simultaneous leaks.

Recently, data-driven methods appeared to overcome the drawbacks of model-based approaches (calibration, WDS complexity, modelling errors). They analyse information gathered from installed sensors to detect and locate leaks. In Romano et al. (2013), multivariate Gaussian mixtures-based graphical models are used for detection; and four different geostatistical techniques are employed to perform localization, by interpolating the probability values of a burst event at every node of the network. Years later, Soldevila et al. (2020) proposed a localization method based on the Kriging interpolation technique to estimate the pressure map, and the study of the pressure residuals.

This article reviews two methods proposed in Romero-Ben et al. (2022) to complementarily solve the BattLeDIM2020 challenge (Vrachimis et al., 2022), achieving the third place in the competition. Their performance is compared to discuss the differences between these two families of methods, review their advantages and drawbacks, and draw conclusions about their suitability and future.

2. METHODOLOGY

The leak localization solution presented to the BattLeDIM2020 competition in Romero-Ben et al. (2022) is composed of two different methods, whose requirements differ due to the philosophy behind their design:

- A data-driven method that requires the structure of the WDS and measurements from pressure sensors.

⋆ The authors want to thank the Spanish national project L-BEST (Ref. PID2020-115905RB-C21) funded by MCIN/ AEI /10.13039/501100011033.

- A model-based strategy using a calibrated model of the network, and pressure and demand information.

2.1 Data-driven methodology

The data-driven approach is based on two phases (its scheme is represented at Fig. 1):

(1) An estimation of the complete state of the network is computed, selecting the hydraulic head associated to each junction as a representative of the node state.
(2) Then, the leak and leak-free states are compared to obtain a set of possible leak candidates.

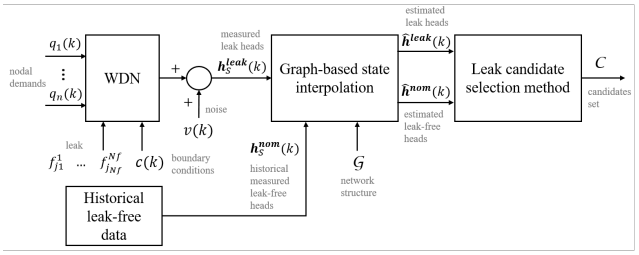

Fig. 1. Data-driven leak localization scheme.

Graph-based state interpolation For the first stage, let us consider the WDS structure to be modeled by the topology of its underlying graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which is simple and directed. The node set $\mathcal{V}$ stands for the junctions of the network, so that the $i-th$ node is denoted as v_i. The edge set $\mathcal{E}$ represents the set of pipes, with edge $e_{ij} = (v_i, v_j)$ linking node v_i with node v_j. Each edge is considered to be characterized by its weight (indicating the importance of the connection between nodes) and its direction (e_{ij} denotes that the edge is traversed from v_i to v_j).

The key idea is to simplify the actual relation between the hydraulic heads of connected nodes in a WDS, which is non-linear (e.g., Hazen-Williams equation), approximating it by an equation that imposes a linear relation, namely:

$$\hat{h}_i = \frac{1}{d_i}\mathbf{w}_i\hat{\mathbf{h}} \tag{1}$$

where $\hat{h}_i$ denotes the state of v_i, i.e., the estimation of h_i; $\mathbf{w}_i \in \mathbb{R}^{1\times|\mathcal{V}|}$ stores the weights of the links between v_i and all the nodes of the graph, derived as the inverse of the pipe length (if $e_{ij} \notin \mathcal{E}$, $w_{ij} = 0$), and $d_i = \sum_{j=1}^{|\mathcal{V}|} w_{ij}$.

Considering Equation (1), the optimal state could be computed through the minimization of following expression:

$$\sum_{i=1}^{n}\left[\hat{h}_i - \frac{1}{d_i}\mathbf{w}_i\hat{\mathbf{h}}\right]^2 = \hat{\mathbf{h}}^T L D^{-2} L\hat{\mathbf{h}} \tag{2}$$

where $L = L^T = D - W$ is the unnormalized Laplacian matrix of $\mathcal{G}$, $W \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the weighted adjacency matrix (composed by the row vectors $\mathbf{w}_i$), and $D \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the degree matrix, which is diagonal with the $i-th$ diagonal value being d_i. In this way, the graph Laplacian and the rest of matrices encode the structural information of the graph during the state estimation.

In order to provide the sensor data, an equality constraint is added to the minimization. If $\mathcal{S} \subseteq \mathcal{V}$ is a set of nodes with a pressure sensor, then $M\hat{\mathbf{h}} = \mathbf{h}_S$, where $M \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{V}|}$ is a matrix whose entry $m_{ij} = 1$ if the $i-th$ sensor is located in the $j-th$ node (and $m_{ij} = 0$ otherwise), and $\mathbf{h}_S \in \mathbb{R}^{|\mathcal{S}|}$ is composed by the hydraulic head values at the sensors.

Finally, the directionality of the edges is imposed through an inequality constraint, i.e., $B\hat{\mathbf{h}} \leq \boldsymbol{\gamma}$, where $B \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$ is the incidence matrix, that assigns $b_{kj} = 1$ if $e_k = e_{ij} \in \mathcal{E}$, $b_{kj} = -1$ if $e_{ji} \in \mathcal{E}$, and $b_{kj} = 0$ otherwise; and $\boldsymbol{\gamma} \in \mathbb{R}^{|\mathcal{E}|}$ is a vector with a value of γ at all its entries, which behaves as a slack in the accomplishment of the direction constraint. The incidence matrix can be approximated by the network structure, by considering the most probable path from the water inlets (sources) and the network inner nodes (sinks).

In conclusion, the optimization problem is formulated as:

$$\begin{aligned} \min_{\hat{\mathbf{h}}} \quad & \frac{1}{2}[\hat{\mathbf{h}}^T L D^{-2} L\hat{\mathbf{h}} + \beta\gamma^2] \\ \text{s.t.} \quad & B\hat{\mathbf{h}} \leq \boldsymbol{\gamma} \\ & \gamma > 0 \\ & M\hat{\mathbf{h}} = \mathbf{h}_S \end{aligned} \tag{3}$$

where γ is considered in the objective function to pursue its minimization, with β controlling its relative importance. Note that this value is a design criteria, although the confidence in the estimation of B must always be considered.

Leak candidate selection method This second stage is based on the comparison between leak and leak-free scenarios to select a set of node candidates, which should be the most probable locations of the actual leak. The interpolation stage must be applied over both scenarios, obtaining the complete leak state vector $\hat{\mathbf{h}}^{leak} \in \mathbb{R}^{|\mathcal{V}|}$ and the complete leak-free/nominal state vector $\hat{\mathbf{h}}^{nom} \in \mathbb{R}^{|\mathcal{V}|}$.

The candidate-selection method can be summarised as:

(1) Both $\hat{\mathbf{h}}^{nom}$ and $\hat{\mathbf{h}}^{leak}$ are used to generate a cloud of 2-D points, representing the x and y-coordinates respectively.
(2) The best-fitting line to the cloud of points is computed. Most of the nodes would remain almost unaffected by the leak, so this line can be used as a prediction of the expected pressure relation when comparing two healthy scenarios.
(3) Thus, the most affected nodes can be retrieved from the most distant points of the cloud to this line. To obtain them, the distance r_i from the point representing node v_i to the line is computed. Note that this distance may be positive or negative, because the point can be above or below the line. Only positive distances are considered, so that the pressure is lower in the leak scenario.
(4) The final set of candidates is selected through a dynamic threshold, represented by the standard deviation of the complete distance vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{V}|}$. The candidates set can be ordered from most to least probable through the information in $\mathbf{r}$, giving the highest probability to the furthest node to the line.

2.2 Model-based methodology

The model-based approach uses a hydraulic model of the WDS that requires a demand estimation module. This module provides the water demands at every node v_i $i = 1, ..., |\mathcal{V}|$ from the total inlet flow measurements and data from automated metered reading devices (AMRs) that can be installed in some nodes of the network. The scheme of this approach is represented in Fig. 2.

The hydraulic simulator computes head estimations at sensor nodes for all the possible leak scenarios, i.e., $\hat{\mathbf{h}}_S^i$ with $i = 1, ..., |\mathcal{V}|$. The comparison between the measured heads, $\mathbf{h}_S$, with the computed leaky heads, $\hat{\mathbf{h}}_S^i$, in a time window N_W after the leak has been detected, is used to obtain the most probable leak location at instant k as:

$$\hat{j}(k) = \arg_i \min \sum_{j=0}^{N_W-1} ||\mathbf{h}_S(k-j) - \hat{\mathbf{h}}_S^i(k-j)||_2 \quad i = 1, ..., |\mathcal{V}| \tag{4}$$

The computation of $\hat{\mathbf{h}}_S^i$ requires an estimation of the leak magnitude $\hat{f}(k)$, that can be computed through the total WDS demand in the time window N_W (Alves et al., 2022).

This method can tackle the problem of multi-leak scenarios if leaks appear at sequential time instants $k_1 < k_2 < .. < k_{N_f}$. In this case, the hydraulic simulator should be updated at every leak localization time k_i with the leak estimation magnitude $\hat{f}(k_i)$ as an extra demand in the node $\hat{j}(k_i)$ obtained in (4).

The performance of this method depends on the hydraulic model accuracy, sensor noise and availability of reliable demand information (Blesa and Pérez, 2018). This third factor is potentially the most critical one because it is not easy to estimate user demands with high accuracy if there are no AMRs installed in some network nodes.

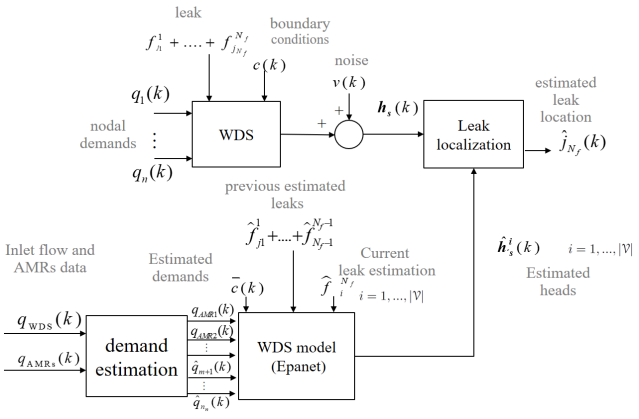

Fig. 2. Model-based leak localization scheme.

3. CASE STUDY AND DISCUSSION

As previously mentioned, the presented methodologies were used to tackle the leak localization task in BattLeDIM2020 competition, based on the L-TOWN benchmark (Vrachimis et al., 2022).

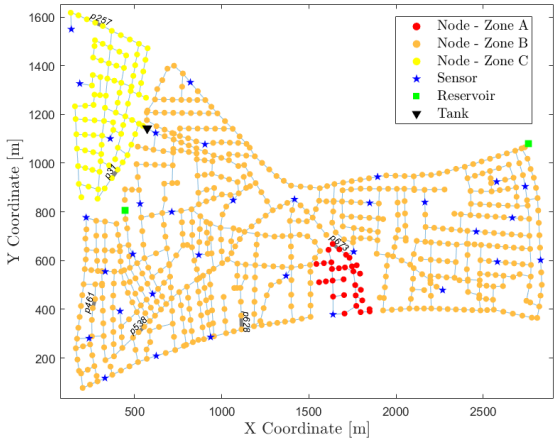

Fig. 3. Structure of the L-TOWN network.

The WDS associated to this benchmark is displayed in Fig. 3. It is divided into three areas:

- Area A: it is the largest area, composed of 659 junctions. This zone is connected to both of the network's inlets, so that it feeds the remaining areas. There are 29 pressure sensors installed throughout this area.
- Area B: it is the smallest zone (31 junctions), connected to Area A through a pressure reduction valve (PRV). There is only 1 pressure meter.
- Area C: this small area (92 junctions) is connected to Area A through a tank. There are only 3 pressure sensors, although there exist 82 AMRs.

In the BattLeDIM2020 competition (Romero-Ben et al., 2022), the data-driven approach (henceforth referred as DD) was applied in Area A, whereas the model-based method (henceforth denoted as MB) was applied to Area B and Area C. Several considerations were taken into account for this choice:

(1) DD requires sufficient pressure data, which is its only source of hydraulic information.
(2) MB's performance is boosted by accurate demand information, whereas lacking this data may hinder the method operation.
(3) Both methods can handle simultaneous leaks.

Then, DD was selected for Area A due to the sufficient pressure sensor density and the lack of demand measurements; whereas MB was applied in Area C due to the limited pressure sensors and high AMRs density; and in Area B due to the lack of pressure sensors.

Thus, this section aims to review the different hypothesis presented above, by applying each one of the methods to the areas where it was not previously considered. To this end, various leaks from the 2018 dataset of BattLeDIM2020 are considered (their locations can be found in Fig. 3).

3.1 Pressure sensor distribution

The existence of sufficient pressure sensors is vital for localization tasks. To highlight the differences between MB and DD, two opposite scenarios are considered:

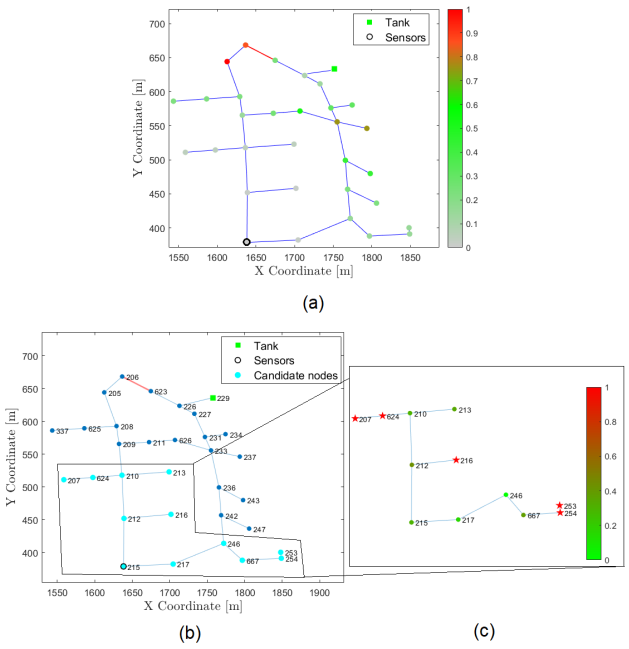

Fig. 4. Localization result for leak 673: (a) Model-based; (b) Data-driven (candidate selection); (c) Data-driven (node-level)

Low sensorization The best zone of the network to study this scenario is Area B, due to the existence of a single pressure meter. A leak event occurred on 2018-03-05 at pipe 673, during until 2018-03-23. The leak localization results yielded by the application of both methods is illustrated in Fig. 4. On the one hand, subfigure 4a shows the result of the application of MB by means of a colour map over the network area: the red colour indicates a high probability of the corresponding nodes to be connected to the leaky pipe, while the green colour represents the opposite. On the other hand, subfigures 4b and 4c compose the result provided by DD: the first image highlights the set of candidates over the network (cyan nodes), whereas the second plot shows a colour map of the leak probabilities. For both methods, a red line over an edge indicates the location of the leak.

The results show how MB outperforms DD, as the former is capable of pinpointing the leak with a negligible error (one node), whereas the latter selects the area near the sensor location, posing as the best candidates (red stars) a set of nodes whose distance to this sensor is similar.

This shows the capability of model-based methods to compensate the side effects of lacking pressure sensors through the knowledge about the network dynamics that the model offers, which is the piece of information that is not accessible to the data-driven approaches. In this extreme case, the existence of a single pressure sensor does not allow sensor redundancy to obtain information of the leak location, hence yielding a solution based on the pressure drop magnitude, resulting in a degraded performance.

Sufficient sensorization This type of scenario is represented by a leak located in Area A, as it is the most sensorized area of the WDS. Specifically, it occurred at

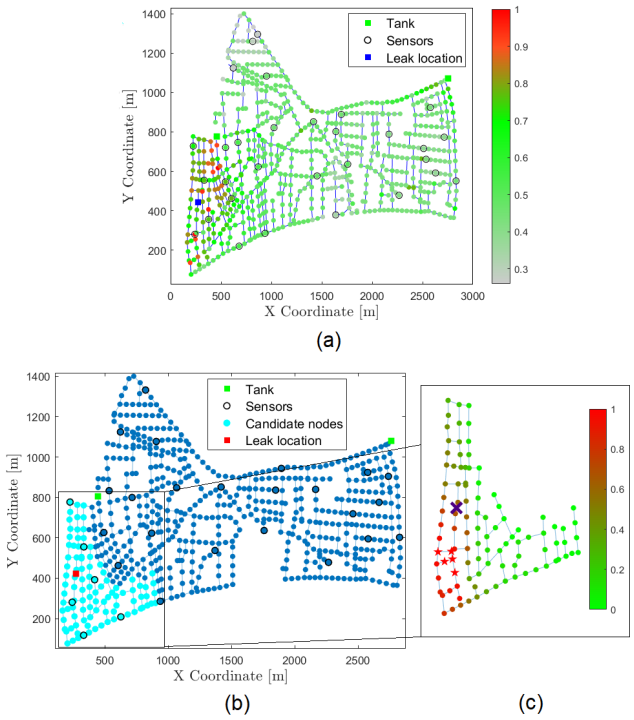

Fig. 5. Localization result for leak 461: (a) Model-based; (b) Data-driven (candidate selection); (c) Data-driven (node-level)

pipe 461 on 2018-01-23, and lasted until 2018-04-02 (this was an incipient leak). The localization results are shown in Fig. 5.

In this case, both MB and DD are able to greatly reduce the leak location area. Unlike the previous scenario, the existence of sensor redundancy makes it possible to bound this area, providing a very accurate result, considering the difference of only a few pipes between the best candidates and the actual leak. Note that a blue cross is added in subfigure (c) to indicate the location of the real leak. This operation is performed in subsequent cases, only when the actual leak is among the highlighted candidates by DD in subfigure (b).

Additionally, it is interesting to highlight how graph-based data-driven methods provide "smooth" solutions considering the transition between high and low probability candidates (see Fig. 5c). This may not occur with model-based methods if they are based on the pressure sensitivities, because they are not explicitly considering structural information (see Fig. 5b, where the best candidates are not connected).

3.2 Demand measurements availability

The implementation of AMRs for demand metering is increasing, although currently water utilities do not typically have access to measured demand information. However, the use of this kind of measurements during localization can be explored. The leak events occurred in Area C are studied to this end, because around a 88% of its nodes are equipped with AMRs. The first leak occurred at pipe 273 on 2018-01-08, and it was not repaired (the results are displayed in Fig. 6). The second leak appeared at pipe 31

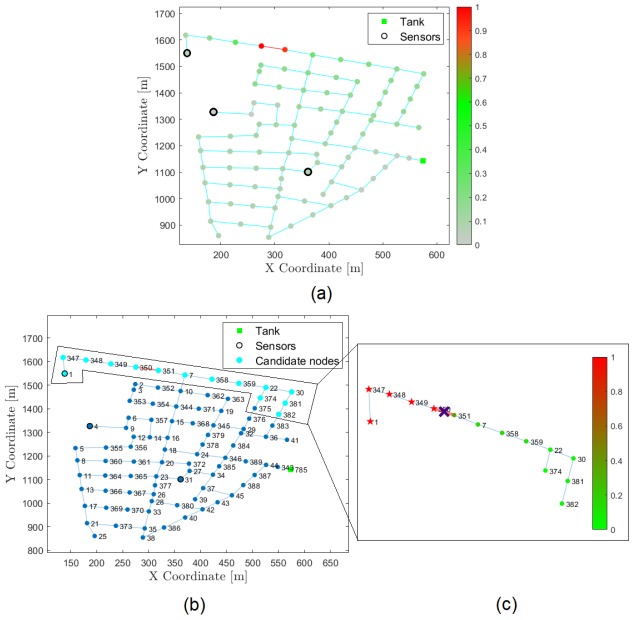

Fig. 6. Localization result for leak 257: (a) Model-based; (b) Data-driven (candidate selection); (c) Data-driven (node-level)

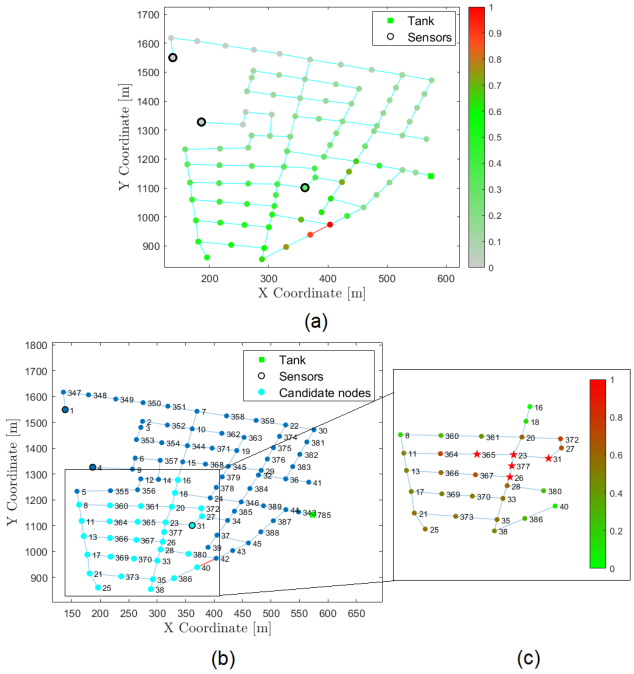

Fig. 7. Localization result for leak 31: (a) Model-based; (b) Data-driven (candidate selection); (c) Data-driven (node-level)

on 2018-06-28, and was repaired on 2018-08-12 (the results are shown in Fig. 7).

On the one hand, the first leak was correctly located by both MB and DD. The former proceeded with a higher degree of accuracy, considering that the leaky pipe is the only one whose nodes are highlighted in red. In the case of DD, the real leak location is among the best candidates, although there are other possible solutions considering the colour map.

On the other hand, the second leak is only correctly located by MB: again, the degree of accuracy is outstanding, considering that the only candidates are the nodes composing the leaky pipe. This also confirms the goodness of MB as a method that solves multi-leak problems, considering that the leak at pipe 257 was not repaired. However, DD fails during the localization operation, not including the leak among the candidates.

These behaviours demonstrate the excellent performance of model-based approaches when precise demand information is available, as well as the problems of data-driven schemes when the amount of pressure sensors is reduced, despite the availability of demand meters.

3.3 Multi-leak solution

The occurrence of simultaneous leaks in a WDS has not been widely considered throughout the literature of leak localization methods. However, this problem is of high interest for water utilities, considering that new leaks may appear while others are already occurring. The performance of MB and DD is compared in this kind of situation.

In the two-leak situation described in the previous section, the existence of precise demand information leaded to MB performing greatly, and the lack of pressure sensors caused DD to perform in a degraded way. Now, the opposite situation is explored, considering leaks occurring in Area A, where there are no AMRs, so that the demand information is much less accurate than in Area C; and there is a higher pressure sensorization density. The next scenario is studied: a leak occurred in pipe 628 on 2018-05-02, and it was not fixed until 2018-05-29. Additionally, a leak appeared at pipe 538 on 2018-05-18. The localization results for these scenarios are displayed in Fig. 8 and Fig. 9.

It may be observed that MB yields the same solution when both leaks are occurring and when only 538 remains active, whereas DD is able to include both leaks in the candidates set, and then marks leak 538 once the previous one is fixed.

This results help us conclude that model-based approaches can struggle to handle multi-leak scenarios if one leak is not correctly localized (considering the necessary model update), while data-driven approaches can yield correct localization results in such situations, because they do not base their operation on the study of previous leaks.

4. CONCLUSIONS

This article presents a comparative analysis of model-based and data-driven leak localization methodologies. The goal of the work is to derive criteria and guidance for the selection of leak localization methods, considering the specific requirements and goals of the problem at hand.

Indeed, utilities may be concerned with different leak size, network size, intervention-crew organization and investment budget for sensor placement, data management and/or model design and calibration. Then, different specific water networks pose different leak localization prob-

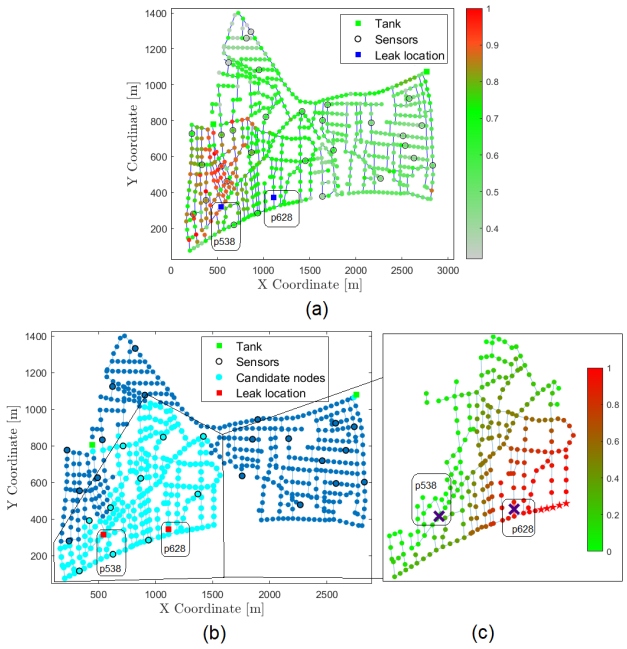

Fig. 8. Localization result for simultaneous leaks 628 and 538: (a) Model-based; (b) Data-driven (candidate selection); (c) Data-driven (node-level)

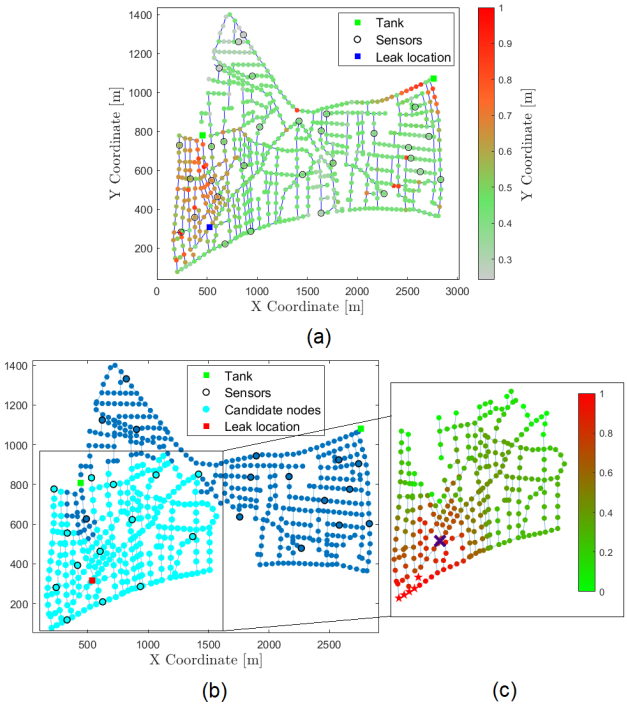

Fig. 9. Localization result for leak 538: (a) Model-based; (b) Data-driven (candidate selection); (c) Data-driven (node-level)

lems, depending on the existence (or not) of a well-calibrated hydraulic model with accurate demand estimation and the availability of sensor data. These are key factors to consider when choosing a valid localization approach. Similarly, model-based and data-driven methods differ in how they cope with multiple-leak or changes in the WDS. All of these aspects are analysed using the benchmark areas of the BattLeDIM2020 challenge.

These results illustrate how model-based approaches are robust even if the number of sensors is reduced, although their performance is conditioned by model calibration and demand data availability. With an appropriate amount and distribution of sensor data and a basic knowledge about network connectivity, data-driven methods can provide valid search areas for leak localization, even if no reliable demand data are available. The comparative analysis provided in this work may prove useful for water network utilities to decide on investments for water loss reduction.

REFERENCES

Alves, D., Blesa, J., Duviella, E., and Rajaoarisoa, L. (2022). Leak detection in water distribution networks based on water demand analysis. *IFAC-PapersOnLine*, 55(6), 679–684. 11th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS 2022.

Blesa, J. and Pérez, R. (2018). Modelling uncertainty for leak localization in water networks. *IFAC-PapersOnLine*, 51(24), 730–735.

Liemberger, R. and Wyatt, A. (2019). Quantifying the global non-revenue water problem. *Water Supply*, 19(3), 831–837.

Perez, R., Sanz, G., Puig, V., Quevedo, J., Escofet, M.A.C., Nejjari, F., Meseguer, J., Cembrano, G., Tur, J.M.M., and Sarrate, R. (2014). Leak localization in water networks: A model-based methodology using pressure sensors applied to a real network in barcelona [applications of control]. *IEEE control systems magazine*, 34(4), 24–36.

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

Romano, M., Kapelan, Z., and Savić, D. (2013). Geostatistical techniques for approximate location of pipe burst events in water distribution systems. *Journal of Hydroinformatics*, 15(3), 634–651.

Romero-Ben, L., Alves, D., Blesa, J., Cembrano, G., Puig, V., and Duviella, E. (2022). Leak localization in water distribution networks using data-driven and model-based approaches. *Journal of Water Resources Planning and Management*, 148(5), 04022016.

Soldevila, A., Blesa, J., Jensen, T.N., Tornil-Sin, S., Fernández-Cantí, R.M., and Puig, V. (2020). Leak localization method for water-distribution networks using a data-driven model and dempster–shafer reasoning. *IEEE Transactions on Control Systems Technology*, 29(3), 937–948.

Steffelbauer, D.B., Deuerlein, J., Gilbert, D., Abraham, E., and Piller, O. (2022). Pressure-leak duality for leak detection and localization in water distribution systems. *Journal of Water Resources Planning and Management*, 148(3), 04021106.

Vrachimis, S.G., Eliades, D.G., Taormina, R., Kapelan, Z., Ostfeld, A., Liu, S., Kyriakou, M., Pavlou, P., Qiu, M., and Polycarpou, M.M. (2022). Battle of the leakage detection and isolation methods. *Journal of Water Resources Planning and Management*, 148(12), 04022068.