2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control in the Water Industry
Valencia (Spain), 18-22 July 2022
doi: https://doi.org/10.4995/WDSA-CCWI2022.2022.14153

AN EXPERIMENTAL STUDY ON EARLY LEAK LOCALIZATION IN DRINKING WATER NETWORKS USING PRESSURE MEASUREMENTS

Yannick Deleuze[1], Arley Nova-Rincon[2], Yves-Marie Batany[3], Teodulo Abril[4], Damien Chenu[5] and Nicolas Roux[6]

[1]Veolia - Scientific & Technological Expertise Department, Maisons-Laffitte, France,
[2]Veolia - Scientific & Technological Expertise Department, Limay France,
[3]Veolia - Scientific & Technological Expertise Department, Limay, France,
[4]Veolia - Technical and Performance Department Latam, Bogotá, D.C., Colombia,
[5]Veolia - Scientific & Technological Expertise Department, Limay, France
[6]Veolia - Scientific & Technological Expertise Department, Aubervilliers, France

[1] yannick.deleuze@veolia.com, [2]arley.nova-rincon@veolia.com, [3]yves-marie.batany@veolia.com,
[4] teodulo.abril@veolia.com, [5] damien.chenu@veolia.com, [6]nicolas.roux@veolia.com

Abstract

Leaks represent a major issue impacting the management and efficiency of Drinking Water Networks (DWN) in cities worldwide. According to the Development Bank of Latin America, by 2018 the losses in DWN range from 40 to 60% in the region. In Europe, the OECD reports a wider range with few losses in cities like Amsterdam (4%) up to 37% in Naples. With this context, some regional policies have emerged like the 2020 european drinking directive "Right2Water", that aims to encourage major suppliers (more than 50000 users), to develop tools to measure and reduce leakages by 2025. Considering this situation, we introduce here a systematic approach for leak management that combines field data, hydraulic models (HM) and machine learning.

Model-based and data driven methods have been of great interest for leak location methodologies in DWN. This research will design energy-efficient and cost-efficient leak localization hotspots in the DWN. The approach is intended for sectorized DWN, equipped with a SCADA system and where a calibrated hydraulic model (e.g. EPANET) is available. This latter serves to evaluate the sensitivity of the system to leaks and identify potential points for pressure measurements in order to optimise the number of installed sensors.

Given a detected leak in the network, a multiclass classifier using pressure data is developed to reduce the inspected pipe length for the leak location. The leakage localization method is implemented combining multiple individual classifiers using ensemble learning methods and a reduced number of decision variables.

The methodology is tested on a real case study from a Colombian site. The method faces challenges in (a) collecting correctly labelled real leak data, and (b) modelling and calibrating hydraulic models. Those challenges are being addressed. The outcome shows that the length of pipes inspected can be reduced by one third with high performance in accuracy with few sensors required (low capital expenditures) and low computational effort (low energy and low operational expenditures).

Keywords
Water distribution systems, Leak zone location, Hydraulic modelling, Mixed model-based/data driven methods.

1 INTRODUCTION

The constant increase of urban populations requires a more efficient management of the water resources. One of the aspects that affects the efficient use of water resources are leaks in DWNs [1], representing the main source of water physical losses in these systems [2]. They can cause significant economic losses in fluid transportation leading to increased reparations and therefore, a potential extra cost for the final user [3]. Since it is a major issue in the performance of DWN, leak detection and management have been a subject of studies with different approaches and important evolutions as detailed in different literature reviews [2, 4, 5].

Leak localization is frequently performed using acoustic methods using hardware such as leak correlators and leak noise loggers and other non-acoustic methods [2]. These methods are accurate, however, it can take extensive labour effort to find a leak, even in small District Metered Areas (DMAs). Leak localization processes can benefit from software-based methods using mathematical modelling in addition to pressure and flow field measurements. Usually, the leak is localised in a zone (a group of network nodes). Hardware-oriented technologies can then be used to identify the exact location of the leak at pipeline level. Hu *et. al*. [6] provide a detailed literature review on model based and data-driven approaches for leak detection and localization. Some leak localization approaches can reduce the area of interest and facilitate traditional DMA techniques to locate leaks. In addition to the possibility of saving large volumes of water, software-based methods also represent financial savings when using only a few pressure sensors.

Mixed model-based/data driven approaches have been developed aiming to predict leak locations. These methods consist in the (i) construction of a hydraulic model, (ii) calibration of the hydraulic model, (iii) leak detection, and (iv) leak localization. After a leak has been detected, it is necessary to localise it within the network. The step (iv) characterises the mixed model-based/data-driven approach using Machine Learning algorithms. Compared to a leak-free situation, a leak results in larger flow in a pipe, with larger head loss, and differences in pressure within the DWN. In a monitored DWN, signatures on pressure data can be used to find leak location [7]. A calibrated hydraulic model can be used to generate synthetic data based on a multitude of demand scenarios. Subsequently the synthetic data can be used to train a classifier. Researchers have used a variety of classification methods. Rojek and Jan Studzinski [8] train Artificial Neural Networks (ANNs) such as Multi-Layer Perceptron (MLP) and Self-Organising Map (SOM). Their goal is to find the exact node where a leak is occurring. De Silva *et. al*. [9] use Support Vector Machines (SVMs) and claim better results compared to Neural Networks. Gupta [10] trains SVMs, ANNs, Random Forest classifiers and Ensemble methods using all the previous methods to improve leak localization predictions. Ares-Milian *et. al.* [1,11] train a SVM classifier to find the leak zone. The authors then apply a model/optimization-based approach to refine and reduce the leakage detection zone. Soldevila and al. [3] present a slightly different approach, where the modelling of the DWN is made and compared to the measured pressures in real time. The comparison is made using residuals. A residual represents the difference between the pressure measured and a reference leak-free scenario pressure from the model. Bayesian classifiers and k-Nearest Neighbour classifiers are trained for the localization task. In this work, the authors overcome the difficulty related with the implicit difference of synthetic data from noisy measured data. The method requires measurements to control all the boundary conditions on the hydraulic model. These constraints can be hard to obtain on most hydraulic networks. Other improvements include the selection of better locations for pressure sensors [12]. From the aforementioned contributions, we can also point out the difficulty to assess the classifiers' performances on real scenarios as collection of leak and leak-free data is hard. To date, model-based leak detection methodologies have not reached the maturity expected by the water industry.

Considering this, this article presents an improved model-based approach for early leak localization in DWNs from pressure measurements. Field data includes days of normal operation without leaks and a set of nine controlled leaks (leak campaigns). On the other hand, a synthetic dataset gathers the result of hundreds of thousands of scenarios generated from a calibrated hydraulic model (EPANET [14]), including demand variations on each node, and leaks of different sizes on every pipe of the network. Classifiers for the leak localization tasks are trained on residuals computed against leak-free scenarios from simulations and known days of normal operation. Then, we propose a domain adaptation technique to correct the differences between synthetic and measured data distributions by searching for a new representation of the data in which both domains (measured and synthetic) are similar. The performance of the trained classifier is evaluated on measured data only. The performance on synthetic data is not of interest. The methodology is tested in a real district metered Area (DMA).

2 METHODS

2.1 Early leak localization methodology overview

The methodology proposed for early leak localization is presented in Figure 1. The proposed method localises leaks in a DWN once a leak has been detected with any procedure available (examples in the literature [5]).

The requirements for the proposed methodology in a DWN application are:

- A hydraulic model (EPANET) of a DWN or a District Metered Area (DMA);
- A set of pressure sensors (installed on the network using sensitivity matrices-based methods[13])
- Inlet and outlet (if applicable) flowmeters to capture the hydraulic information;
- An IT system for data acquisition and processing;
- A calibration procedure for the hydraulic model.

The classifying task for early leak localization is computed in two main stages. First the network is partitioned into a set of candidate leak zones using sensitivity matrices [13] considering the topological relationships between the network nodes. Leak zones correspond to groups of pipes with similar hydraulic response to leaks (see section 2.3). Then a trained multiclass classifier is used to estimate a potential leak zone (see sections 2.4).

2.2 Training and testing sets

Machine learning algorithms require an important number of training examples. Nevertheless, obtaining real-world DWN leak data is near-to-impossible, since it would imply collecting data for leaks in all pipes multiple times. The lack of historical data is a great challenge. We will have at our disposal only a limited set of samples. In addition, these records are, most of the time, not properly labelled. For example, start and end days of the leak and/or location might not be properly logged.

The training database will be constructed from simulated scenarios of variable demands. Simulations are run using EPANET Programmer's Toolkit. On the other hand, the test database will be created with measured pressure data. Leaks are created by opening purge valves or fire hydrants. The advantage of synthetic data is that it is properly labelled and balanced for machine learning applications. Although DWN models remain only an approximation (even if calibrated), the use of synthetic data is commonly accepted in the field of machine learning [15].

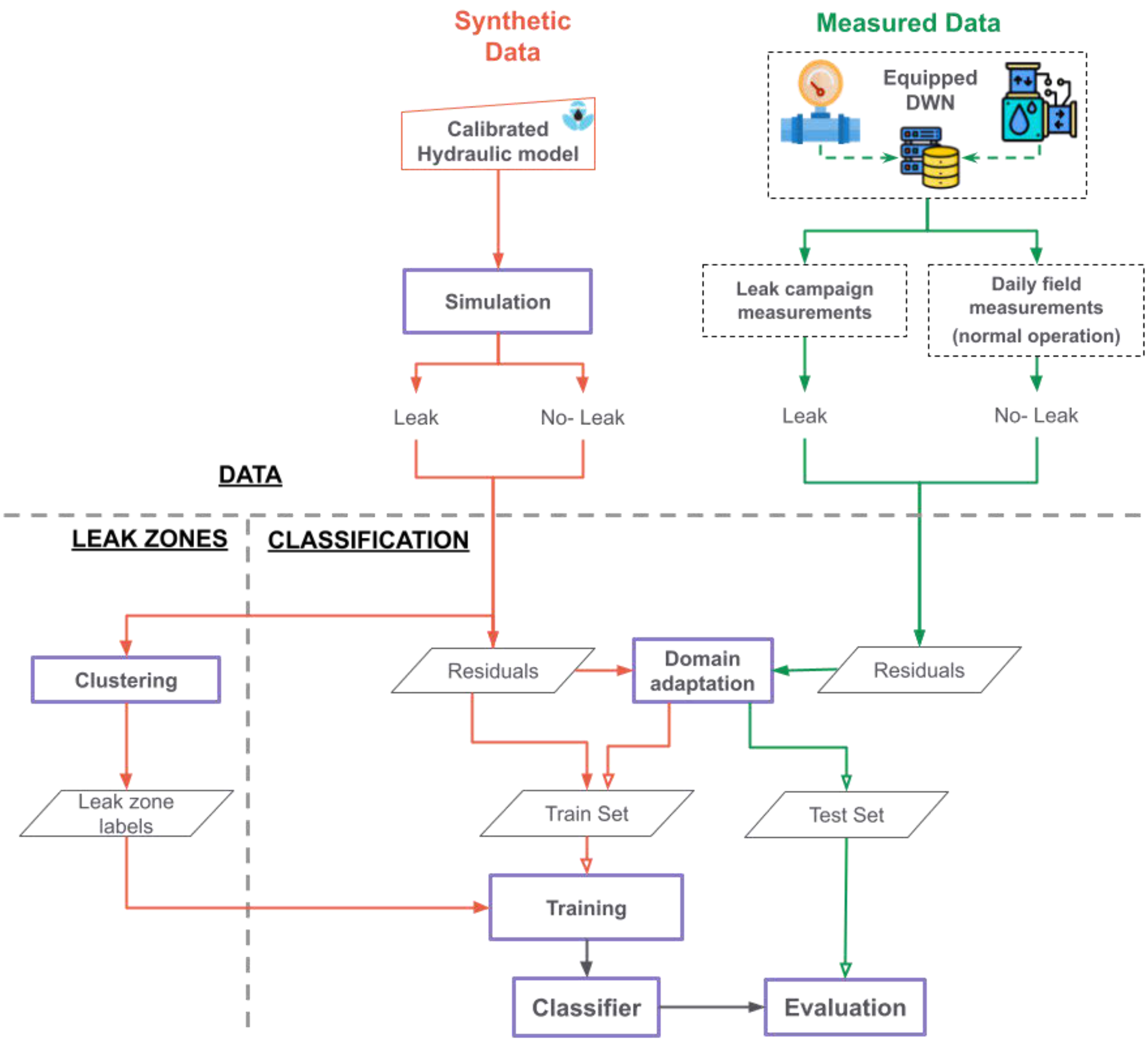

Figure 1. Methodology overview for early leak localization

2.3 Leak Zone labels

Using sensitivity matrices methodologies [13], multiple groups of pipes are defined from the position of sensors and their sensitivity to leaks. A sensitivity leak analysis is done by simulating leaks with different flow rates at all possible pipes within the DWN and calculating the corresponding sensitivity in flow and pressure at each point by comparing with the base-case scenario (i.e. no leaks in the network). The two main limitations to directly use these groups are that they are often groups of one pipe and groups of multiple pipes can include pipes not located in the same part of the DMA.

The spectral clustering method [16] has proven itself to be effective in many different situations. The network of pipes is considered as a mathematical graph with a set of edges and a set of vertices. The mapping from the hydraulic model to the graph model is as follows: pipes (and valves) correspond to edges and junctions (such as pipe intersections, water sources) correspond to nodes. Spectral clustering is applied considering only the distance between nodes.

The search zones can be constructed from groups of pipes defined by hydraulic sensitivity to leaks and groups of pipes defined from the spectral clustering with an ensemble method [17]. Finally, labels for the multiclass classifier are defined by matching every sample with the zone that contains the actual leak node.

2.4 Classifier and input variables

In this study, we propose the use of the Gradient Boosting Classifier [18], that is a set of machine learning algorithms which include several weaker models that are combined together to get higher predictive output. Gradient Boosting algorithms are popular due to their ability to classify datasets effectively. Finding the leak zone where a leak occurred can be framed as a multiclass classification task. The Scikit-learn [20] project provides a set of machine learning tools that can be used for the multiclass classification task.

In order to remove long-term seasonality, the classification tasks proposed will not be carried out directly on the pressure and flow raw variables. We use residuals to remove a short-term mean behaviour of the network from the measured signal. To do so, we take the average of M days to classify and subtract the mean over N past days labelled as having normal behaviour (no leaks, no outliers). A process to identify days with normal conditions is thus essential. The residuals are defined by Equation 1:

$$X_{res}(t) = \frac{\bar{X}_M(t) - \bar{X}_N(t)}{\sigma(X_N(t))} \tag{1}$$

where:

- $\bar{X}_M(t) = \frac{1}{M}\sum_{m=1}^{M} X_m(t)$ is the mean value of a sensor at time t over the M consecutive days to classify;
- $\bar{X}_N(t) = \frac{1}{N}\sum_{n=1}^{N} X_n(t)$ is the mean value of a sensor at time t over N previous days that exhibit a normal behaviour; and
- $\sigma(X_N(t))$ is the standard deviation of values of a sensor at time t over N previous days that exhibit a normal behaviour.

Parameters N and M are chosen depending on the network hydraulics and precisions of the sensors. Repetition helps isolate signals from noise. Note that for synthetic data, N and M do not represent a number of consecutive days. Simulations are unordered as each simulation with the EPANET software is independent of others. Days can be selected randomly for the simulation scenarios. The times t can be either chosen during the day, a window during night time, on a precise time of the day such as the time at which the minimal night flow is observed.

One issue we face when using synthetic data is that our training and testing data are different. Training data is drawn from simulation from the EPANET model and thus has a bias (from modelling) and variability different from the testing data drawn from measurements. The purpose of transfer learning and domain adaptation methods [19] is to handle this common issue encountered in machine learning. In the domain adaptation setting, one considers, on one hand a source domain from which a large sample of labelled data is available. And on the other hand, a target domain from which no labelled data are available. One refers to unsupervised domain adaptation. The goal is to build an estimator on the target domain by leveraging information from the source domain.

CORAL (CORrelation ALignment) [21] is a feature based domain adaptation method which minimises domain shift by aligning the second-order statistics of source and target distributions. Feature-based methods consist of searching for common features which have similar behaviour with respect to the classification task on source and target domain. A new feature representation (often called encoded feature space) is built with a projecting application which aims to correct the difference between source and target distributions. The CORAL method transforms source features to minimise the Frobenius norm between the correlation matrix of the input target data and the one of the transformed input source data. CORAL can use only labelled source and unlabeled target data. The ADAPT Python package [22] provides tools that can be used for domain adaptation. The classification task is then trained in this encoded feature space instead of the initial input space.

2.5 Performance metrics

2.5.1 Accuracy score

The accuracy score is defined in Equation 2 as the percentage of leak samples that were correctly located by the classifier for N samples:

$$accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{L_i}(\hat{L}_i) \quad (2)$$

where L_i is the actual leak zone for sample i, $\hat{L}_i$ is the predicted leak zone for sample i, and $\mathbb{I}$ the indicator function.

2.5.2 Top-k accuracy score

For the multiclass classification problem, another metric of interest is the top-k accuracy score. The leak zone prediction is considered correct as long as the true leak zone is associated with one of the k highest leak zones predicted by the classifier. For N samples, the top-k accuracy is defined in Equation 3 as:

$$topk\ accuracy = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} \mathbb{I}_{L_i}(\hat{L}_{i,j}) \quad (3)$$

where L_i is the actual leak zone for sample i, $\hat{L}_{i,j}$ is the j-th largest score of the predicted leak zones for the sample i, and $\mathbb{I}$ the indicator function. This metric is interesting for leak localization as it can measure the leak localization performance along a path to search the DWN.

2.5.3 Matthews correlation coefficient

The Matthews correlation coefficient (MCC) is used in machine learning as a measure of the quality of classification. The MCC ranges between 1 and -1. A coefficient of 1 represents a perfect prediction, 0 an average random prediction, and a negative score is linked to an inverse prediction. The MCC is defined for N samples by Equation 4:

$$Matthews\ correlation\ coefficient = \frac{c \times N - \sum_{k=1}^{N_L} p_k \times \gamma_k}{\sqrt{(N^2 - \sum_{k=1}^{N_L} {p_k}^2) \times (N^2 - \sum_{k=1}^{N_L} {\gamma_k}^2)}} \quad (4)$$

where γ_k is the number of times a leak truly occurred in leak zone k, p_k is the number of times a leak was predicted in leak zone k, c the count of leak zones correctly predicted, and N_L the number of leak zones. Having an unbalanced and few number of samples of each leak zone, this measure provides a metric on leak zones not being predicted by randomness.

2.5.4 Leak Search Zone Size

On top of machine learning metrics, we propose to look at the expected gain in terms of reduced leak search zone. The expected leak search zone size is defined in Equation 5 as the percentage of the total DWN linear length to explore to truly find a leak:

$$Leak\ Search\ Zone\ Size = \frac{D - \sum_{k=1}^{N_Z} (\frac{c_k}{N} \times \sum_{j \neq k}^{N_Z} d_j)}{D} \times 100 \quad (5)$$

where c_k is the number of times a leak is correctly predicted in leak zone k, d_j is the pipe length of leak zone j in km, D the total pipe length in the DWN , N the number of samples, and N_Z the number of leak zones. A correct prediction by the classifier would avoid the need to inspect the rest of the network.

3 CASE STUDY AND RESULTS

3.1 Case study

The methodology was developed and implemented on a Colombian DWN operated by Veolia Colombia. The DWN is located in a residential sector, and the INP model provided by the operator accounts for 288 nodes (users with individual demands), 303 pipes for a total length of 13.5 km approximately, a reservoir representing the unique source of water for the whole sector, and 3 pressure reduction valves (PRV). In 2019 the sector reported a total water loss of 18% (leaks +apparent losses). Figure 2 shows topology of the sector of analysis.

The set of data used in this work, contains the measured values of 12 pressure sensors and 1 flowmeter, whose locations are also detailed in Figure 2. Data was collected for a year (September 2020 to August 2021), including mostly normal operation measurements, as well as data of 2 leak campaigns, over 5 different points (1 to 5 in Figure 2). The field data frequency is 1 measurement every 15 minutes.

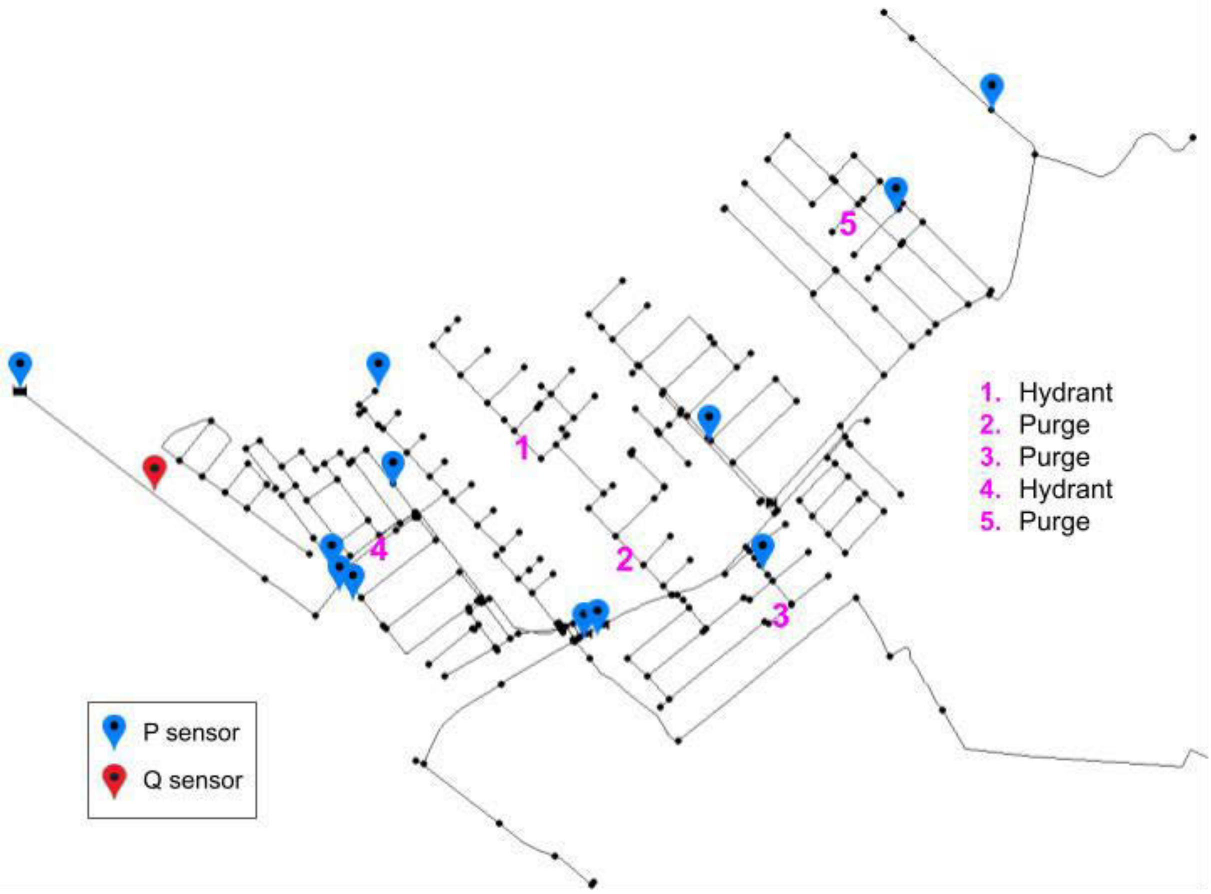

Figure 2. Location of the sensors and controlled leaks in the case study sector

The leak campaigns consisted of a total of 9 leak configurations (Table 1), each one with a duration of 26 hours. The first leak campaign was launched in February 2021 (configurations A and B) and the second one in June 2021. Three of the configurations considered simultaneous leaks in two of the five points. For the sake of consistency, at least two days of normal operation were considered between each consecutive configuration to avoid any possible effect of the precedent controlled leak.

Table 1. Leak campaign configurations

Config	Location	Max flow rate [m^3/h]	Config	Location	Max flow rate (m^3/h)
A	2	1	G	3	5
B	5	3		5	2
C	2	5	H	2	5
D	3	3		4	2
E	1	2	I	5	3
F	2	5		1	2

3.2 Experimental results

When implementing the methodology in the case study, we tested the division of the sector into different numbers of clusters ranging from 2 to 8. Figure 3 details the resulting zones for 2, 4, 6 and 8 leak zones. Residuals have been computed with M=1, N=5, at time t when the minimum night flow is observed. The training dataset contains 15000 simulations. Performance of the proposed method is evaluated on the leak campaign test dataset.

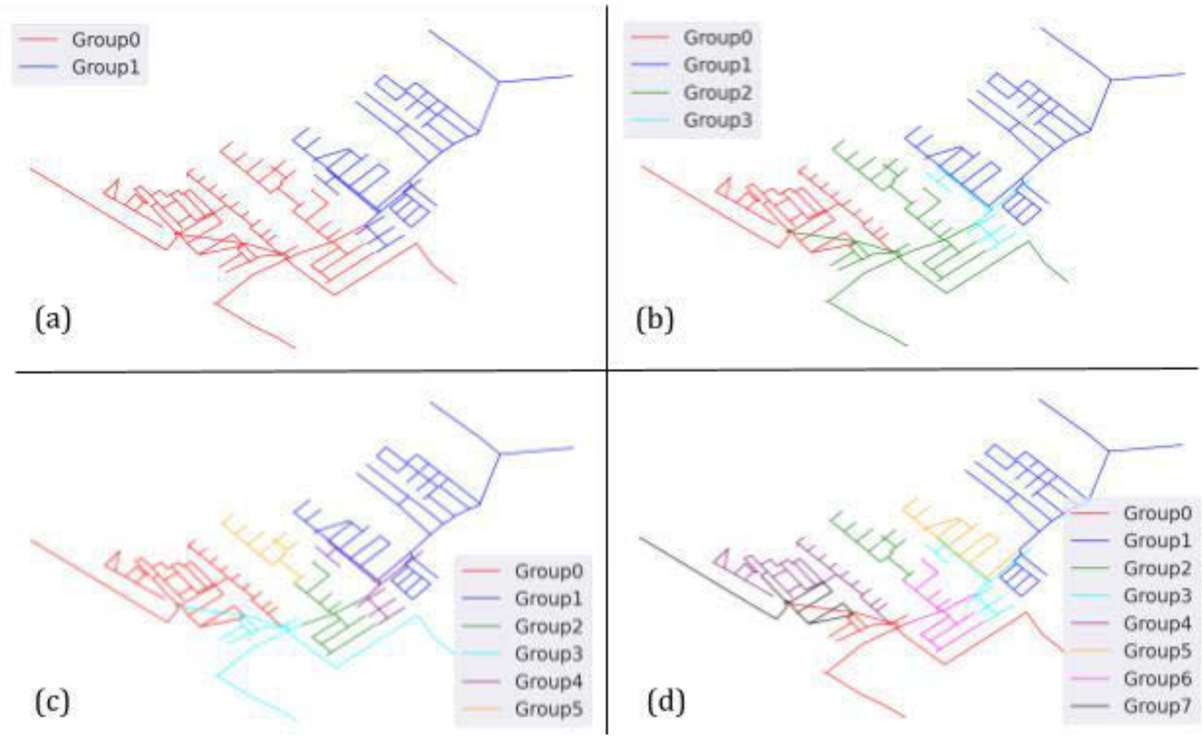

Figure 3. Leak zones for different numbers of clusters [(a) 2, (b) 4, (c) 6, (d) 8]

Figure 4 and Figure 5 present respectively the accuracy score and the top-2 accuracy score for the different number of leak zones to search within. As expected, the lesser number of zones, the higher the accuracy of the classifier. Using domain adaptation to compute a feature encoded space before training the classifier seems to improve the accuracy of the classifier. Without domain adaptation, the classifier is trained directly on the pressure residuals. With domain adaptation, the 2-zone classifier reports an accuracy of 78%. The top-2 accuracy score remains over 80% up to 4 zones.

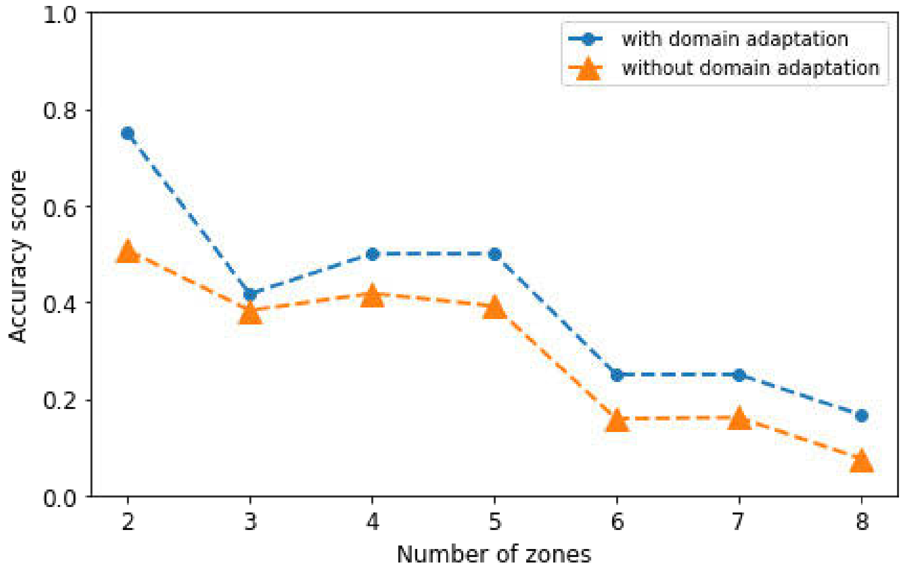

Figure 4. Accuracy score by number of zones

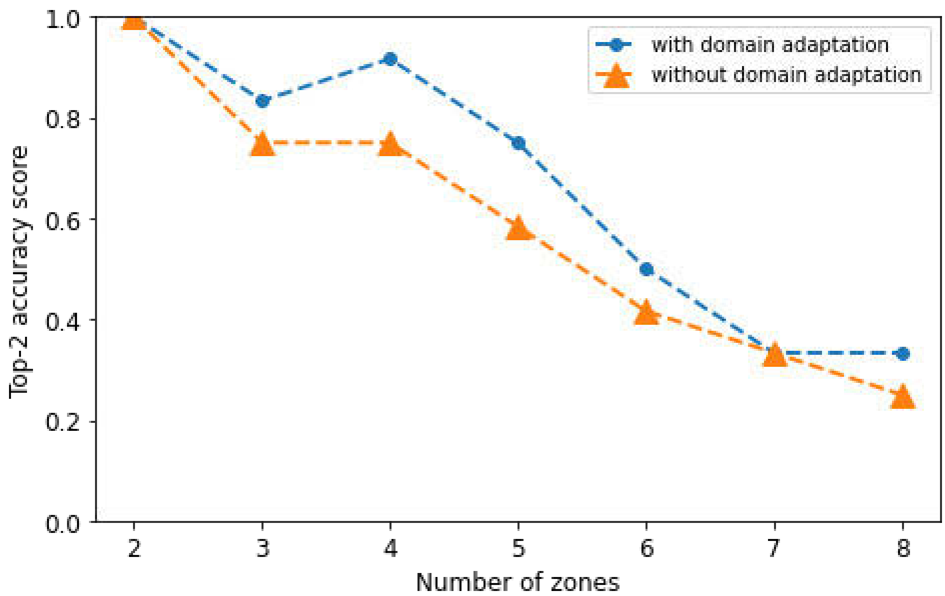

Figure 5. Top-2 Accuracy score by number of zones

The Matthews correlation coefficient score is depicted in Figure 6. The MCC score can help to identify the ineffectiveness of the classifier. For 6 and more zones, the score is below 0.2. In this case, the classifier is very close to a random guess classifier. For 2 to 5 zones, the classifier can be considered better than a random guess classifier.

On a field perspective of the method's performance, the percentage of the expected leak zone search is presented in Figure 7. The 2-zones classifier yields a leak search zone of just above 60% of the total DWN pipe length. With the proposed method, the length of pipe inspected is expected to be reduced by one third. The 4-zones classifier also yields promising gains with a leak search zone just below 70% of the total DWN.

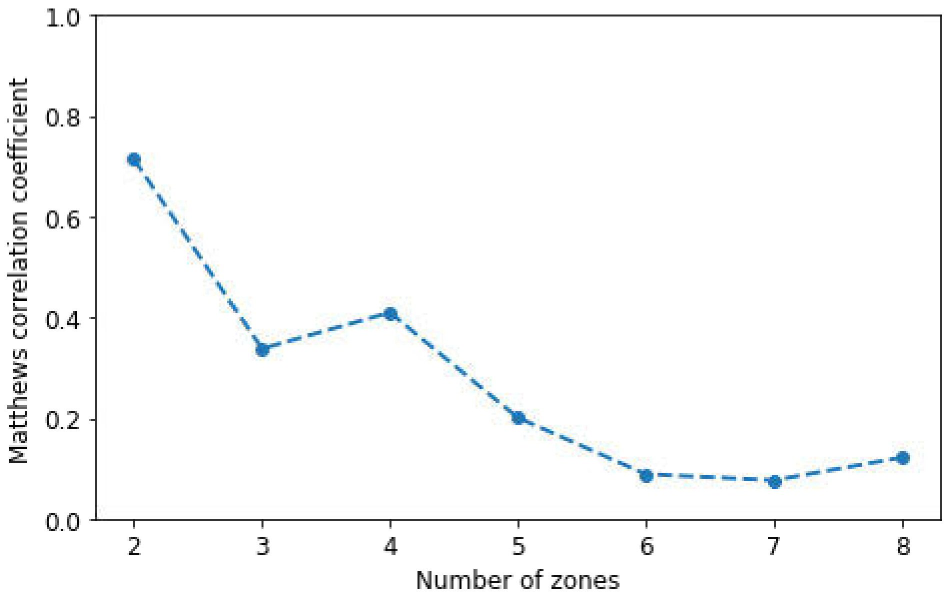

Figure 6. Matthews correlation coefficient by number of zones

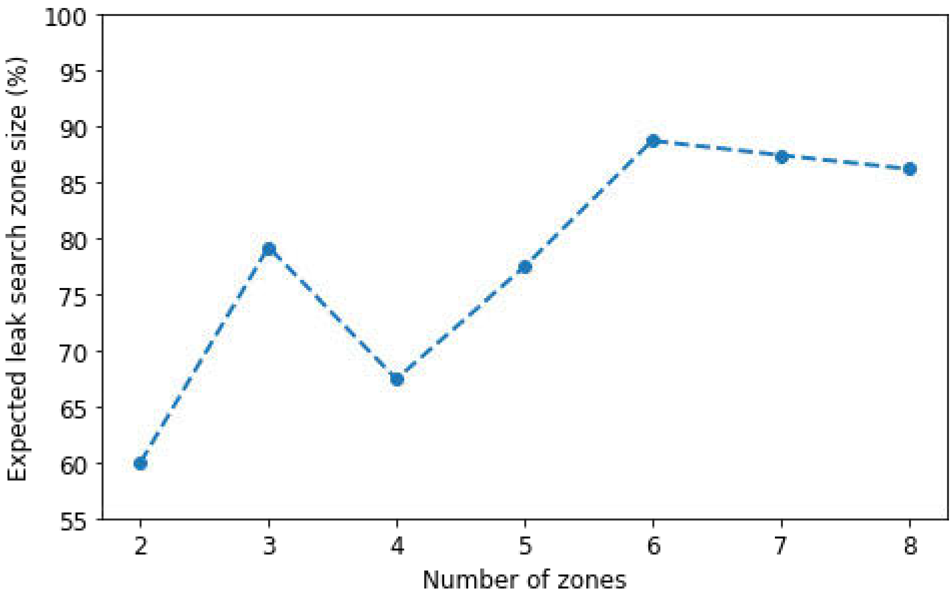

Figure 7. Expected leak search zone size in percentage of the total DWN pipe length by number of zones

4 CONCLUSIONS

Reducing costs and time searching leakages in DWNs is known to be a difficult task for the water utility industry. The presented work demonstrates the benefits of using a hybrid model-based and data-driven approach for early leak localization. Dividing the DWN into two or four zones doesn't seem like the target, but shows success in contributing to early leak localization by reducing the leak search space. This low cost solution requires only pressure sensors on top of common inlet and outlet flowmeters, and it relies on more frugal (less data, parameters, and computational resources required) algorithms than neural networks. The field experiment considers 12 sensors, but with few zones, the number of required sensors could be lower.

Proper data collection of network hydraulic behaviour for data-driven approaches is a challenging task. Historical data are either non-existent or not properly labelled. To overcome this issue, our experimental approach focuses on evaluating the performance only on measured data from a designed leak campaign. This strategy allows us to go further than using synthetic data or toy DWN models. To reduce the differences between synthetic data samples and the measured data samples, a domain adaptation approach was implemented. Moreover, the proposed residuals don't require such high fidelity simulation of the DWN. The current limitation lies in the identification of previous normal operation days to be able to compute a mean pressure profile for each of the sensors.

In conclusion, the proposed methodology allows to improve early leak detection performance, while meeting operational requirements. The field experiments show that the total pipe length required for inspection in the DMA can be reduced by one third with a significant gain of performance in accuracy, while requiring a limited number of pressure sensors installed on the network.

5 REFERENCES

[1] M. Quiñones-Grueiro, M. Ares Milián, M. Sánchez Rivero, A. J. Silva Neto, and O. Llanes-Santiago, "Robust leak localization in water distribution networks using computational intelligence," Neurocomputing, vol. 438, pp. 195–208, May 2021, doi: 10.1016/j.neucom.2020.04.159.

[2] A. Gupta and K. D. Kulat, "A Selective Literature Review on Leak Management Techniques for Water Distribution System," Water Resour Manage, vol. 32, no. 10, pp. 3247–3269, 2018, doi: 10.1007/s11269-018-1985-6.

[3] A. Soldevila, J. Blesa, S. Tornil-Sin, E. Duviella, R. M. Fernandez-Canti, and V. Puig, "Leak localization in water distribution networks using a mixed model-based/data-driven approach," Control Engineering Practice, vol. 55, pp. 162–173, Oct. 2016, doi: 10.1016/j.conengprac.2016.07.006.

[4] T. K. Chan, C. S. Chin, and X. Zhong, "Review of Current Technologies and Proposed Intelligent Methodologies for Water Distributed Network Leakage Detection," IEEE Access, vol. 6, pp. 78846–78867, 2018, doi: 10.1109/ACCESS.2018.2885444.

[5] S. El-Zahab and T. Zayed, "Leak detection in water distribution networks: an introductory overview," Smart Water, vol. 4, no. 1, p. 5, Jun. 2019, doi: 10.1186/s40713-019-0017-x.

[6] Z. Hu, B. Chen, W. Chen, D. Tan, and D. Shen, "Review of model-based and data-driven approaches for leak detection and location in water distribution systems," Water Supply, vol. 21, no. 7, pp. 3282–3306, Apr. 2021, doi: 10.2166/ws.2021.101.

[7] S. Sophocleous, D. Savić, and Z. Kapelan, "Leak Localization in a Real Water Distribution Network Based on Search-Space Reduction," Journal of Water Resources Planning and Management, vol. 145, no. 7, p. 04019024, Jul. 2019, doi: 10.1061/(ASCE)WR.1943-5452.0001079.

[8] I. Rojek and J. Studzinski, "Detection and Localization of Water Leaks in Water Nets Supported by an ICT System with Artificial Intelligence Methods as a Way Forward for Smart Cities," Sustainability, vol. 11, no. 2, p. 518, Jan. 2019, doi: 10.3390/su11020518.

[9] D. De Silva, J. Mashford, and S. Burn, "Computer aided leak location and sizing in pipe networks (second edition)," Urban Water Security Research Alliance, Technical Report EP112167, Apr. 2011.[Online]. Available: http://hdl.handle.net/102.100.100/104745?index=1

[10] G. Gupta, "Monitoring Water Distribution Network using Machine Learning," KTH Royal Institute of Technology, Stockholm, Sweden, 2017. Accessed: Apr. 12, 2022. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-221832.

[11] M. J. Ares-Milián, M. Quiñones-Grueiro, C. Verde, and O. Llanes-Santiago, "A Leak Zone Location Approach in Water Distribution Networks Combining Data-Driven and Model-Based Methods," Water, vol. 13, no. 20, p. 2924, Jan. 2021, doi: 10.3390/w13202924.

[12] A. Soldevila, J. Blesa, R. M. Fernandez-Canti, S. Tornil-Sin, and V. Puig, "Data-Driven Approach for Leak Localization in Water Distribution Networks Using Pressure Sensors and Spatial Interpolation," Water, vol. 11, no. 7, p. 1500, Jul. 2019, doi: 10.3390/w11071500.

[13] 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, Oct. 2011, doi: 10.1016/j.conengprac.2011.06.004.

[14] L. A. Rossman, “The EPANET Programmer’s Toolkit for Analysis of Water Distribution Systems,” pp. 1–10, Apr. 2012, doi: 10.1061/40430(1999)39.

[15] M. Hittmeir, A. Ekelhart, and R. Mayer, “On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks,” in Proceedings of the 14th International Conference on Availability, Reliability and Security, New York, NY, USA, août 2019, pp. 1–6. doi: 10.1145/3339252.3339281.

[16] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888–905, Aug. 2000, doi: 10.1109/34.868688.

[17] L. Rokach, “Ensemble-based classifiers,” Artif Intell Rev, vol. 33, no. 1, pp. 1–39, Feb. 2010, doi: 10.1007/s10462-009-9124-7.

[18] J. Friedman, “Greedy function approximation: A gradient boosting machine.,” 2001, doi: 10.1214/AOS/1013203451.

[19] F. Zhuang et al., “A Comprehensive Survey on Transfer Learning,” Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[20] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Apr. 15, 2022. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[21] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, février 2016, pp. 2058–2065.

[22] A. de Mathelin, F. Deheeger, G. Richard, M. Mougeot, and N. Vayatis, “ADAPT : Awesome Domain Adaptation Python Toolbox,” arXiv:2107.03049 [cs], Jul. 2021, Accessed: Apr. 15, 2022. [Online]. Available: http://arxiv.org/abs/2107.03049