A flexible, leak crew focused localization model using a maximum coverage search area algorithm

Brett Snider[1], Gareth Lewis[1], Albert Chen[1], Lydia Vamvakeridou[1,2] and Dragan Savić[1,2]

[1] Centre for Water Systems, University of Exeter, Exeter, EX4 4QF, UK
[2] KWR Research Institute, Nieuwegein, Netherlands

b.snider@exeter.ac

Abstract. Buried watermains are deteriorating and pipe failure is increasing in many cities. In response, advanced leak location models have been developed to help identify where a leak is occurring – which allows utilities to react quickly to pipe bursts and reduce the impact of the leak. This paper develops a new leak location model that is designed to identify optimal search areas for leak crews using a random forest classification model and the maximum coverage location problem algorithm. The model, when compared with other machine learning and clustering localization predictions, reduces the search space by over 35%, allowing utilities to confirm leak location and mitigate its impact more efficiently. The new model is also highly customizable, able to adjust the number of search areas and search size quickly and easily to meet leak crews' requirements.

1. Introduction

Beneath all cities lay a network of underground watermains that deliver drinking water to their residents and businesses. Many of these buried watermains have experienced significant deterioration, resulting in a major increase in pipe breaks in recent years [1]. The impacts from these pipe breaks can be severe, resulting in a loss of service, possible contamination, and impact to nearby infrastructure. Therefore, it is a priority for water utilities to locate and repair these leaks as soon as possible to help mitigate their impact. However, locating a leak is usually a very difficult task since the length of buried watermains that must be checked can be extensively long (for example London, England's water distribution system contains over 20,000 kms of buried watermains [2].

In recent years a variety of leak location models have been developed that can help predict where a leak is likely occurring within the network. These tools aim to improve leak crews' response time, by narrowing their focus, allowing them to locate the leak sooner and mitigate the leak's impact. However, from an operations point of view, these models still leave a lot of room for improvement. First, many of the location models developed attempt to pinpoint an exact location of the leak. However, leak crews are still required to search in-person to confirm the location of the leak. If the leak is not at the predicted location, which is often the case, crews are left wondering where to search next. Second, none of the existing models are customized to fit the utilities leak crews' preferences, such as search area size and/or number of search areas, which may be highly dependable on staff availability for instance.

The model presented in this paper addresses these concerns by designing a leak location model that identifies optimal search areas for leak crews to confirm leak location. The predicted search area size and number of search areas can be customized to fit the utilities demands at that time of leak detection. The model developed applies the Maximum Coverage Location Problem algorithm (MCLP) [3] to a node probability prediction from a Random Forest Machine Learning (RF) classifier model. This new leak location modelling approach is compared with other machine learning and clustering approaches to leak localization to evaluate and ultimately demonstrate its effectiveness.

The remainder of this paper analyses the effectiveness of the proposed leak location model by first providing a quick overview of the literature regarding leak location models. A detailed methodology outlines the model's development. A case study is then used to highlight the impacts of the model's customizable parameters and provide a comparison between other leak location models.

2. Leak Location Models

A digital transition is occurring within the water industry. Many water utilities have begun to realize the numerous benefits associated with digitization of their assets and water systems. Remote sensors are becoming more popular and leak management models more common. In recent years a large increase in leak location models has been developed. The leak location models that have been developed use a variety of methods to locate leaks within a water distribution system and can be broadly classified into three main approaches: transient based, data driven, or model driven methods [4].

Transient based methods use high frequency pressure sensors within the water distribution system to detect transient pressure waves resulting from a leak occurrence. By analysing this pressure waves and identifying the signal variation between sensors, transient based models attempt to identify the location of the leak within the WDS. A variety of transient based models have been developed and described in detail by various literature reviews [5], [6]. However, transient based models require high frequency sensors, which can be very costly and may not be applicable for real-time application or for use in large complex water networks [5], [6].

Data driven leak location models do not require any specific knowledge regarding the water distribution system. Instead, signal processing and statistical analyses are used to detect and locate leaks [7]. Most data driven leak models only focus on leak detection, however, it is possible to develop a leak location model using data driven techniques[4], [8]. Overall, data driven methods require significant amount of historic sensor readings to develop an accurate prediction model and may be inaccurate when water network experiences seasonal or festive variation that is common within water networks [7]. Due to these drawbacks, the focus of this paper is on the development of model driven leak location methods.

The most popular approach to leak location prediction is via model driven methods. These methods rely on hydraulic models of the water distribution system (WDS), comparing simulation reads to recorded sensor measurements to help identify where a leak is occurring within the network. Model driven methods are able to be calibrated without requiring significant historic sensor readings and have been shown to have a high degree of accuracy when hydraulic models are well calibrated [4]. The various model driven methods used to detect leak location include sensitivity matrix-based approaches, mixed model-based approaches, optimization-calibration approaches, and error-domain falsification-based approaches [4]. Instead of focusing on the specific modelling-based method, this paper concentrates on the output these models generate, as this has a direct impact on how the model is utilized by the utility/leak crews.

2.1. Leak Location Prediction Output

Typically, the leak location models predict a specific geographical point, either through hydraulic model node classification, or x-y coordinates using a regression analysis [9]. Using these methods, leak crews are sent to the predicted location and if the leak is not there, they are left without any guidance on where to search next.

A few leak location models have been designed to predict a zone within the network that is likely to contain the leak. The leak zone prediction models have been developed in two distinct ways;

through classification of predefined leak zones / district metered areas [9], [10], or through iterative cluster / graph theory partitioning [11], [12]. By splitting up the network into predefined leak zones, these location models predict which zone the leak is occurring in. Utility leak crews are then sent to search that leak zone and confirm the leak location. These leak zone approaches provide boundaries to help guide the leak crews. However, they do not place limits on leak zone size, resulting in some leak zones being very large and would take a long time to search.

Lacking within the literature is a leak location model specifically developed for leak search crews. A location model designed for leak crews should identify search areas, based on the crews preferred search radius and number of search locations that contain the highest probability of leak. By developing this type of model, leak crews would be able to locate and respond to the leak quicker, minimizing the overall impact of the leak on the customers, utility, and nearby infrastructure. This paper sets out to fill this literature gap by applying the MCLP algorithm to a nodal leak location prediction from a RF classification model. This approach allows the specific search area radius and number of search areas to be defined by the utility, while outputting the largest probability of including the leak within these parameters.

3. Methodology

The following section describes the MCLP leak location prediction model developed as well as the development of four other leading leak location prediction used evaluate the effectiveness of MCLP modelling approach.

The overall development of the leak location model involves three main steps: developing a training dataset, training the RF classification model, and applying the MCLP algorithm to predict leak locations. The overall development of the model is highlighted in Figure 1 with inputs and outputs depicted as well.

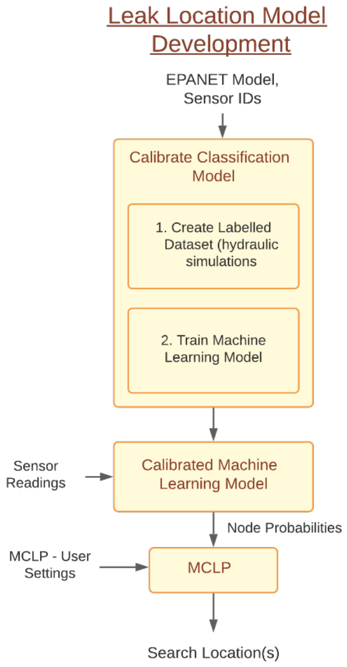
Leak Location Model Development
EPANET Model, Sensor IDs
Calibrate Classification Model
1. Create Labelled Dataset (hydraulic simulations
2. Train Machine Learning Model
Sensor Readings
Calibrated Machine Learning Model
Node Probabilities
MCLP - User Settings
MCLP
Search Location(s)

Figure 1: MCLP leak location model development.

3.1. *Creating a labelled dataset*

The first step in developing the MCLP location model is to create a labelled dataset to train the RF classification model. The dataset is generated by performing several leak simulations for every node within the hydraulic model, storing only the readings associated with sensor values (i.e., flow and/or pressure) and the location, or node, of the simulated leak. The simulated sensor values are converted into residuals (or z-score) based on a no-leak scenario, as highlighted in Equation 1 below:

$$z_{i,t} = \left(x_{i,t} - xavg_{i,t}\right) / \sigma_{i,t} \tag{1}$$

where *z* represents the calculated residual, *x* the sensor measurement (flow or pressure), *i* the specific sensor, *t* the reported time step; $xavg_{i,t}$ and $\sigma_{i,t}$ are the average and standard deviation for sensor readings calculated from the non-leak scenario for that particular sensor and time period. For each leak simulation a six-hour window of residuals are stored, starting six hours before detection, and ending when the burst is detected (which is randomly chosen to occur between 1-4 hours after burst occurrence).

3.2. Training a random forest classification model

A random forest machine learning classifier is trained using the labelled dataset. Here each leak node represents a class, and the input variables are the residuals for each simulated sensor reading during the six-hour detection window. A random forest machine learning algorithm is chosen due to its high performance reported in other leak location models [9], [13] and its ease of hyper-parameter tuning and robustness against data outliers [14].

The hyper-parameters are tuned using a grid-search approach for tree-depth and number of trees. The tuning parameters are optimized to maximize classification accuracy using a hold-out validation dataset. The RF machine learning classifier is then used to predict the probability score for each class (or in this case node within the hydraulic model). The predicted probability for each node is used as an input to the Maximum Coverage Solution Problem in order to identify the optimal search area(s).

3.3. Maximum coverage location problem

The Maximum Coverage Location Problem (MCLP) is a type of geographical location set covering solution that was developed in 1974 by Church & Revelle [3] but has garnered significant attention in recent years as large GIS datasets have been developed and used to solve unique problems [15].

To identify the optimal search areas for leak detection crews, the MCLP is applied to the output from the RF classification model. Since the RF classification model predicts the probability of the leak occurring at each node within the network, the MCLP can be used to group these results geographically, using the node's geographic coordinates, and locate the search areas that maximizes the predicted likelihood of containing the leak event.

The MCLP algorithm adapted to leak location is as follows:

1. $Maximize\ \ z = \sum_{i \in I} a_i y_i$,
2. $Subject\,to: \ y_i \leq \sum_{j \in N_i} x_j \ \ i \in I$,
3. $\sum_{j \in J} x_j = p$,
4. $0 \leq y_i \leq 1, \ i \in I$,
5. $x_j \in \{0,1\}, \ j \in J$,

where:

i, I	the index and set of EPANET nodes (i.e. possible anomaly locations);
j, J	the index and set of search area centroids;
a_i	the predicted probability at node *i*;
$d_{i,j}$	the shortest distance from node *i* to search area centroid *j*;
S	the search area radius;
N_i	$\{ j \mid d_{i,j} \leq S \}$ = the nodes *j* that are within a distance of *S* to node *i*;
p	the number of search areas to be determined;

x_j a binary variable that equals one when the search centroid is located at the *j*-th node;

y_i a binary variable which equals one if node *i* is within one or more search areas.

To solve the MCLP a list of possible search area centroid locations is required. However, search area centroid locations are typically not defined by a utility / leak detection crew since crews are often flexible enough to go to any location throughout the network. Therefore, a grid of possible search areas that cover the entire network is created using the hydraulic model's nodes with maximum and minimum latitude and longitude coordinates to create a grid boundary. A grid of possible search area centroids within this boundary is then generated using a set distance between grid points.

In addition to the grid of search areas centroids, preferred search area radius and number of search areas must be provided to solve the MCLP. Various grid spacing, search area radiuses, and number of search areas are compared and discussed in the results section.

4. Case study

An EPANET hydraulic model representing a UK water distribution system is used to evaluate the impact grid size, search radius and number of search areas have on the MCLP approach as well as the model's overall effectiveness.

A schematic of the case study's EPANET water network is presented in Figure 2 below.

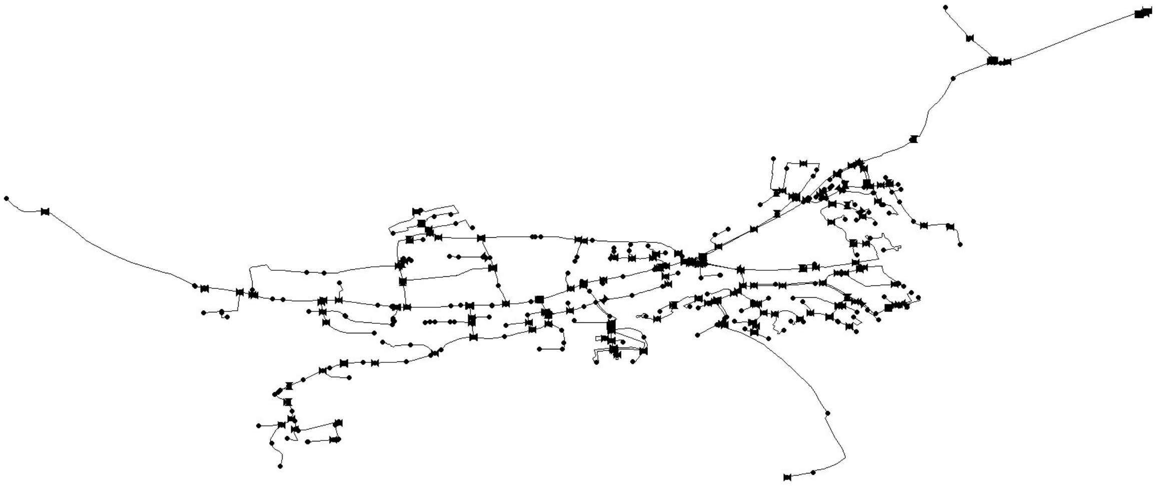

Figure 2: Example EPANET schematic.

The EPANET network is a gravity fed network, supplied by a single reservoir. The hydraulic model is set-up to run with 15-minute extended period analysis and is assumed to have a weekly cycle (i.e., steady-state is reached with a weekly pattern of flows, pressures and demands). The overall attributes of the model are highlighted in **Table 1** below:

Table 1: Greater Torrington hydraulic model attributes.

Network attribute	Value
# of Nodes	1005
# of Pipe Segments	783
Total length of Pipes (km)	25.25
Avg. Pipe Diameter (mm)	118
Avg. Daily Demand (LPS)	13.43
Avg. Pressure (m)	48.0
Geographic Area (km^2)	12

Hypothetical pressure and flow sensors are modelled within the EPANET at thirteen locations throughout the network. Sensor readings are obtained by performing a hydraulic analysis using the EPANET hydraulic model, with the readings corresponding to the link or node id of the sensor for every 15-minute time increment.

4.1. Leak simulations

To build a comprehensive training dataset, six leaks are simulated at every node with random start times. The leaks are simulated using EPANET's emitter equation:

$$Q = C\,P^{\gamma} \tag{2}$$

where Q is the leakage flow, C is the emitter coefficient, P is the nodal pressure and γ is the emitter exponent. The emitter coefficient is randomly selected between 0.5 to 3 and the emitter exponent set to 0.9.

To reflect realistic scenarios, noise is added to user demand and sensor readings. Demand noise is added to each node by multiplying a noise factor to the original demand at that time step. The noise factor is selected from a normal distribution with an assumed standard deviation of 0.25. Noise is also added to the simulated sensor readings to reflect measurement inaccuracy that is typical with all sensors. The sensor noise is added to the sensor reading and is assumed to follow a normal distribution with standard deviation of 0.5.

The sensor readings simulated during the leak scenario are converted into residuals using the residual equation (Equation 1). A six-hour window of residuals is stored for each simulation, starting six hours before leak detection awareness. For testing purposes, the leak detection time is randomly chosen to occur between 1 to 4 hours after burst occurrence.

5. Results

5.1. Random forest machine learning classification model

To tune the RF hyper-parameters, a grid search methodology was employed, and results were tested using a validation dataset which included leak scenarios for 100 randomly selected leak nodes. The tuning parameters assessed include:

- Tree depth: [100, 200, 300, 400, 500, 600];
- Number of trees: [25, 50, 100, 150, 200, 250, 300].

The results from this analysis suggest the optimal tuning parameters were a tree depth of 250 with 300 decision trees developed. The overall accuracy of the classification model tested on an independent test dataset was 7.5%. These results suggest the model can identify important variations in sensor data to detect the leak node location, since the model's accuracy is roughly 75 times greater than chance. However, this is still a relatively low accuracy from a leak detection perspective, highlighting the importance of identifying leak search areas (that are more likely to contain the leak), instead of specific location predictions.

5.2. MCLP sensitivity analysis

A sensitivity analysis is performed for each of the MCLP parameters, which include grid size, search radius and number of search areas. All sensitivity analyses were performed using an independent test dataset set of 100 randomly generated leak scenarios.

Table 2 reflects the impact of grid spacing on predicting leaks located by MCLP approach assuming two search areas, each with a search radius of 50m. Table 2 highlights the trade-off between computation time and accuracy of locating leaks within the MCLP search area by adjusting the grid spacing of search area centroids. As the grid spacing becomes smaller, more search areas must be evaluated by the MCLP algorithm. This improves the likelihood of identifying the optimal search location and its likelihood of containing the leak but results in greater computation time. It is up to the requirements of the utility to determine the grid spacing that works best for them (weighing the trade-off between computational requirements and leak location accuracy).

Table 2: Impact of grid spacing on predicted leaks located by MCLP.

Grid spacing (m)	Computation time (sec)	Predicted probability covered	Leaks located
10	278	40.5%	64%
15	131	39.6%	64%
25	51	38.8%	63%
40	23	36.1%	60%
60	11	34.4%	60%
80	8	31.2%	52%

The second parameter analysed is the search radius. This analysis is performed using the 100-leak test set for two search areas, with 25 m grid spacing between centroids and various search radii.

As shown in Figure 3, as the search radius increases the likelihood of the search area containing the leak increases as well. However, a large search radius would require leak crews to search a larger area before locating the leak. Therefore, choosing a search radius will depend on a utility's own preference for accuracy and total search area (i.e., a trade-off between resolution and accuracy).

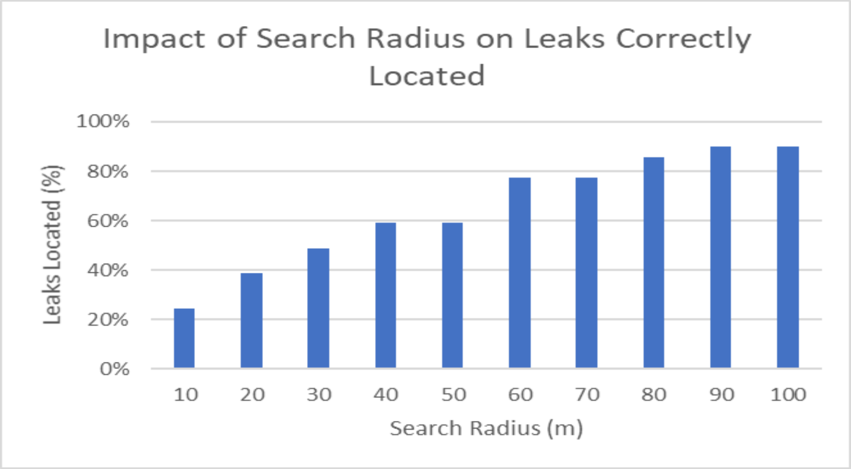

Figure 3: Search radius versus leaks located within search areas.

Lastly, the MCLP approach was analysed using a various number of search areas. A utility with several search crews may prefer several search areas to be identified to increase the speed of which the leak location is confirmed. The following table highlights how the number of search areas impact the accuracy of identifying the leak location using the 100-leak test set, with 25m grid search and 50m search radius. As the number of search areas increase the accuracy of detecting the leak increases as well. However, again a trade-off between total search area and accuracy must be made by the utility.

Table 3: Impact of search areas on anomaly detection.

Number of search areas	Predicted probability covered	Leaks detected	Total search area (m^2)
1	25%	45%	5,027
2	40%	59%	10,053
3	51%	65%	15,080
4	59%	73%	20,106

Overall, the sensitivity analysis performed highlights one of the major strengths of the MCLP approach – that is its customizability. The MCLP leak location model can be adjusted to fit the utilities specific computation and/or labour resources.

5.3. *Model comparison*

The Random Forest MCLP model developed is compared with other machine learning and clustering localization models. Specifically, the Random Forest MCLP approach detailed in this paper is assessed using one search area (MCLP-1 SA) and two search area (MCLP – 2 SA), expanding the search radius until 50% of the simulated leaks within the test set are accurately identified within the search area(s). The required search areas are then compared with three other leading location models:

1. A classification random forest model (Classification - 1 SA) - where the node with the highest predicted probability being chosen as the centroid for the search area. A single circular search area is expanded out from this centroid until 50% of the leaks from the test dataset are within the search areas.
2. A regression based random forest model (Regression - 1 SA) - where the model predicts the latitude and longitude of the search area. A single circular search area is expanded outwards from the predicted latitude and longitude until 50% of the leaks from the test dataset are within the search areas.
3. An agglomerative hierarchical clustering model (Clustering - 2 SA) – where a random forest model classification model is trained to predict probabilities of the leak occurring at each node for each simulation. Using this output, an agglomerative hierarchical clustering algorithm is performed using a max linkage parameter set to the desired search diameter [16]. The search diameter is expanded until two clusters include 50% of the leaks from the test dataset.

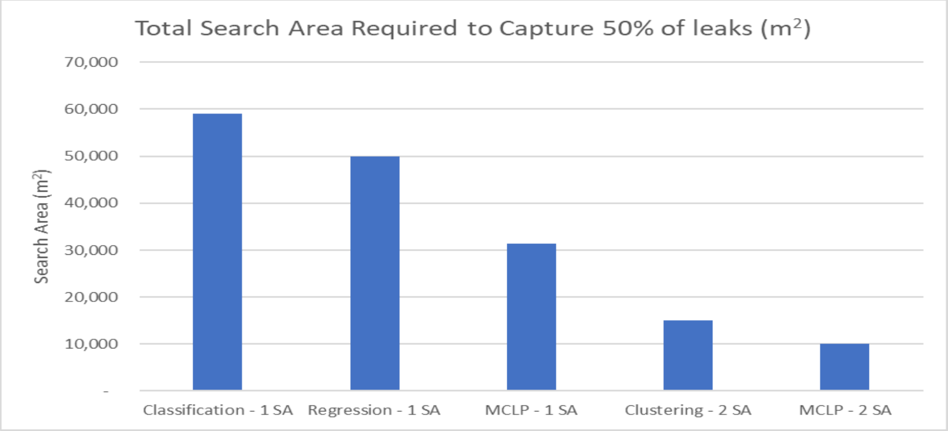

Figure 4: Comparing search areas for various location methods.

Figure 4 identifies the geographical search area required by each model to accurately locate 50% of the simulated leaks within the test dataset. The results indicate that the Random Forest MCLP approach requires the smallest single search area (35% less than leading regression-based model) and the smallest two search area (34% less than the clustering model). Overall, these results indicate the Leak location model developed in this report is effective at reducing the search space required to locate the leak and outperforms the other leak location models assessed.

6. Conclusion

Watermains throughout the world are deteriorating and pipe breaks appear to be increasing [1]. Leak location models that are able to predict the location of leaks can be a very useful tool to help utilities respond to pipe bursts quickly and mitigate their impact. For these tools to be effective they must be designed with the end-user in mind, which in this case would be leak crews that are tasked with confirming the leak location and begin leak mitigation response.

This paper develops a new leak location model, specifically designed to reduce the leak search area required by leak crews to locate and mitigate the leak. Specifically, a MCLP approach is applied to the output from a RF machine learning classification to predict ideal search area(s) for leak crews to locate a leak.

The new model is compared with other machine learning and clustering localization models and the results indicate the MCLP approach significantly reduces the search space required for leak crews to locate the leak. By reducing the search space, the leak crews will be able to locate the leak quicker and mitigate the impacts to the customer and the utility. The MCLP approach also has the important benefit of being highly customizable, being able to adjust the size of the search area, and/or the number of search areas to meet leak crews' preferences.

Overall, the MCLP model is a promising approach to leak location. Future research will continue to advance this applicability of this model. Specifically, developing a heuristic approach to identifying the ideal search area centroid for the MCLP approach could improve the accuracy and computation requirement for the leak localization model. Also, a more in-depth analysis into the model's sensitivity to hydraulic models' inaccuracies and sensor reading inaccuracies would help identify level of calibration and sensor accuracy is needed before the model is implemented within a water distribution system. The current approach described in this paper assumed a set level of noise/inaccuracy with sensors and hydraulic model. Adjusting this inaccuracy level until the model fails to predict the location would allow a utility to understand the accuracy needed before implementing the leak localization model [17]. Also worthwhile to investigate would be the impact of background losses, or small ongoing leaks, on the localization model's accuracy, since all water distribution systems have some level of background water loss. Analyzing the impact these background losses can have on the leak localization accuracy will allow water utilities to determine whether this approach is viable for their particular system.

Acknowledgment

The work presented in this paper was funded by the ongoing European Commission H2020 Fiware4Water (GA 821036) and aqua3S (GA 832876) projects.

References

[1] Folkman S 2018 Water main break rates in the USA and Canada: A comprehensive study *Mech. Aerosp. Eng. Fac. Publ.* **174** https://digitalcommons.usu.edu/mae_facpub/174

[2] Thames Water 2021 *Your Water Services* https://www.mendeley.com/reference-manager/library/all-references

[3] Church R and Revelle C 1974 The maximal covering location problem *Pap. Reg. Sci.* **32**(1) 101–118

[4] Hu Z, Tan D, Chen B, Chen W and Shen D 2021 Review of model-based and data-driven approaches for leak detection and location in water distribution systems *Water Supply* **21**(7) 3282–3306 doi: 10.2166/WS.2021.101

[5] Adedeji KB, Y Hamam, BT Abe and AM Abu-Mahfouz 2017 Towards achieving a reliable leakage detection and localization algorithm for application in water piping networks: An overview *IEEE Access* **5** 20272–85 doi: 10.1109/ACCESS.2017.2752802

[6] Colombo AF, P Lee and BW Karney 2009 A selective literature review of transient-based leak detection methods *J. Hydro-Environment Res.* **2** 212–227 doi: 10.1016/j.jher.2009.02.003

[7] Chan TK, CS Chin and X Zhong 2019 Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection *IEEE Access* **6** 78846–67 doi: 10.1109/ACCESS.2018.2885444

[8] Murvay PS and I Silea 2012 A survey on gas leak detection and localization techniques *J. Loss Prev. Process Ind.* **25**(6) 966–973 doi: 10.1016/J.JLP.2012.05.010

[9] Ares-Milián MJ, M Quiñones-Grueiro, CC Corona and O Llanes-Santiago 2021 Clustering-based partitioning of water distribution networks for leak zone location *Lect. Notes Comput. Sci.* (including *Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics*) **12702** LNCS 340–350 doi: 10.1007/978-3-030-93420-0_32

[10] Zhang Q et al 2016 Leakage zone identification in large-scale water distribution systems using multiclass support vector machines *J. Water Resour. Plan. Manag.* **142**(11) 04016042 doi: 10.1061/(ASCE)WR.1943-5452.0000661

[11] Chen J, X Feng and S Xiao 2020 An iterative method for leakage zone identification in water distribution networks based on machine learning *Structural Health Monitoring* **20**(4) 1938–56 https://doi.org/10.1177/1475921720950470

[12] Shekofteh M, M Jalili Ghazizadeh and J Yazdi 2020 A methodology for leak detection in water distribution networks using graph theory and artificial neural network *Urban Water J.* **17**(6) 525–533 https://doi.org/10.1080/1573062X.2020.1797832

[13] Lučin I, B Lučin, Z Čarija and A Sikirica 2021 Data-driven leak localization in urban water distribution networks using big data for random forest classifier *Mathematics* **9**(6) doi: 10.3390/MATH9060672

[14] Breiman L 2001 Random Forests *Machine Learning* **45**(1) 5–32 https://doi.org/10.1023/A:1010933404324

[15] Elhabyan R, W Shi and M St-Hilaire 2019 Coverage protocols for wireless sensor networks: Review and future directions *J. Commun. Networks* **21**(1) 45–60 doi: 10.1109/JCN.2019.000005

[16] scikit-learn 2021 “2.3. Clustering – scikit-learn 1.0.2 documentation” https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering (accessed Apr. 03, 2022)

[17] Sophocleous S, Savić D and Kapelan Z 2019 Leak localization in a real water distribution network based on search-space reduction *J. Water Resour. Plan. Manag.* **145**(7) 04019024 doi: 10.1061/(ASCE)WR.1943-5452.0001079.