MethodsX 12 (2024) 102612

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

GA-Sense: Sensor placement strategy for detecting leaks in water distribution networks based on time series flow and genetic algorithm

Ary Mazharuddin Shiddiqi[a,*], Choiru Za'in[b], Artya Lathifah[c], Tohari Ahmad[a], Diana Purwitasari[a]

[a] Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia
[b] Department of Computer Science and Information Technology, La Trobe University, Australia
[c] Department of Industrial and Information Management, National Cheng Kung University, Taiwan

ARTICLE INFO

Method name:
GA-Sense

Keywords:
Genetic algorithm
Weighted fitness function
Sensor placement strategy
Time series data
Leak detection and localization

ABSTRACT

The detection of leaks in time series flow systems is crucial for efficient and integrated industrial processes. This is especially true when daily demand patterns differ, as this results in fluctuations in the snapshots of water consumption that are commonly used as the basis for placing sensors to detect leaks. This paper introduces a novel method in which the genetic algorithm (GA) is applied to find optimal sensor locations and to enhance the accuracy of leak detection in time series flow data. The method consists of two steps. Firstly, the GA is used to identify the optimal sensor locations using a specific fitness function that accounts for flow patterns, system topology, and leak characteristics. The novelty of the proposed method lies in the weighting scheme of the fitness function, which takes into consideration the frequency of events and the magnitude of leaks at potential locations. Secondly, the selected sensor locations are integrated with an advanced time series data analysis to locate leaks. In this technique, the most consistently performing locations are dynamically selected over time, allowing the model to adapt to varying conditions to maintain optimal sensor placement. Experiments were conducted on a simulated time series flow system with known leak scenarios to evaluate the performance of the proposed method. The results demonstrated the superiority of our GA-based sensor placement strategy in terms of leak detection accuracy and efficiency compared to other methods.

- We developed a model called GA-Sense for sensor placement strategy by considering flow patterns to maximize leak detection and localization capabilities.
- GA-Sense uses time series data to find strategic sensor locations to identify abnormal flow patterns indicative of leaks.
- This approach enhances the accuracy and efficiency of leak detection and localization compared to alternative methods.

[*] Corresponding author.
E-mail address: ary.shiddiqi@if.its.ac.id (A.M. Shiddiqi).

https://doi.org/10.1016/j.mex.2024.102612
Received 18 November 2023; Accepted 10 February 2024
Available online 10 February 2024

Specifications Table

Subject area:	Computer Science
More specific subject area:	Optimal Sensor Placement in Water Distribution Networks
Name of your method:	GA-Sense
Name and reference of original method:	Casillas MV, Puig V, Garza-Castañón LE, Rosich A. Optimal sensor placement for leak location in water distribution networks using genetic algorithms. Sensors (Basel). 2013 Nov 4;13(11):14,984–5005. doi: 10.3390/s131114984. PMID: 24,193,099; PMCID: PMC3871061.
Resource availability:	N/A

Method details

Introduction

Leaks have given rise to longstanding challenges in water distribution networks (WDNs) due to the non-linear nature of WDNs, which makes leak detection challenging [1]. Various research efforts, including water hammer analysis and flow and pressure observation, have been undertaken to enhance leak detection in water systems. The strategic placement of sensors is crucial to improve the effectiveness of leak detection. Flow-based sensor placement relies on the identification of susceptible locations, such as near pump discharges or along primary flow routes, where flow fluctuations may indicate potential leaks. Ideally, placing sensors on every pipe would be optimal; however, due to practical constraints and the complexity of these networks, an effective sensor placement strategy is necessary. Traditional approaches often rely on data from one-time leakage scenarios, and neglect the importance of time series data, which are important in identifying the most effective locations for sensor placement under various conditions. One method of optimizing sensor placement involves the use of a genetic algorithm [2–6], while some researchers [7] have employed the lean graph method for the detection of pipe leaks.

Recent advancements in sensor-based analysis and artificial intelligence (AI)-based prediction methods have significantly impacted sensor placement strategies for leak detection and localization. The integration of model-based leak detection techniques means that fewer pressure measurements need to be taken, while improving the overall performance of leak localization in WDNs [2]. For example, AI-based acoustic leak detection devices can employ machine learning to differentiate between leakage and normal flow sounds within pipelines, to optimize sensor placement and enhance the accuracy of leak detection [8]. Although AI-based methods have given promising results, such methods may struggle with noisy data, and require extensive training data to enable them to distinguish leakage patterns accurately. A summary of the strengths and weaknesses of the sensor placement methods discussed here is shown in Table 1.

In this study, the genetic algorithm (GA) is used, due to its ability to discover optimal or nearly optimal solutions to a wide range of search and optimization problems. This technique mimics the principles of natural selection to identify the best solution [9]. The effectiveness of the genetic algorithm relies on the fitness function, which plays a pivotal role in its success: a well-designed fitness function can help the algorithm reach its full potential [10]. One approach is to modify the fitness function to determine a good scheme for sensor placement on pipes, based on the frequency of occurrence of leaks [11].

The primary goal of this research is to develop an effective sensor placement strategy using the GA to localize leaks. A key focus of this study is to investigate the impact of designing a good fitness function on the performance of the GA. This research aims to identify the best sensor locations on pipes by analyzing the occurrence of leaks. The effectiveness of sensor placement is evaluated by assessing the performance of the GA in terms of localizing leaks. This work makes two key contributions. The first is the development of an innovative GA fitness function in which a weighting scheme is used to select sensor locations by accounting for the frequency of events and the magnitude of leaks. An improved fitness function can help to prevent the GA from producing local optima solutions. The second contribution is the dynamic identification of the most reliable sensor locations over time, ensuring optimal sensor placement while adapting to fluctuating conditions.

Table 1
Summary of sensor placement methodologies (strengths and weaknesses in terms of leak detection and localization).

Strategy	Strengths	Weaknesses
Genetic algorithm (GA)	Optimizes sensor locations for coverage and cost Can be adapted to various objectives and constraints	Computationally intensive for large networks May converge to local optima rather than global solutions
Machine learning-based methods	Can predictively model complex leak scenarios Gives enhanced accuracy with continuous learning and adaptation	Requires large datasets for model training May be affected by noisy data and false positives
Graph-based methods	Utilizes a network topology for efficient sensor distribution Often simpler and more scalable than optimization algorithms	Less effective for dynamically changing systems May not account for non-topological factors affecting leaks

Literature review

Water distribution networks

A WDN transports drinking water from a centralized treatment plant or good water source to the consumer's tap. This complex system includes various types of components, including pipes, pumps, valves, tanks, reservoirs, meters, fittings, and other hydraulic equipment [12]. Such systems can range in size from small plants serving groups of as few as 25 people to urban networks catering to millions [13]. As the distribution system constitutes the majority of the physical infrastructure for water supply, it poses significant operational and public health management challenges. In addition to ensuring high-quality water for consumption, a key aspect of the distribution system is maintaining adequate water flow. To meet this requirement, a water distribution system uses vertical pipes, overhead tanks, reservoir storage, and larger pipes to maintain consistent flow and capacity, especially over long distances, such as between treatment plants and consumers.

However, a major concern for these networks is water loss through leaks, which can represent up to 30% of the total water supplied in many countries [14]. This significant loss not only has financial implications, but also represents a waste of valuable natural resources. Hence, the development of effective leak detection and localization systems is crucial to conserve significant quantities of water and financial resources.

Time series data

The diagnosis of conditions or issues in water distribution networks is more reliable and convincing when based on time series monitoring data, rather than relying solely on single-point data [14]. Time series data consist of observations of the same element or variable at various points in time or over different periods [14,15]. In water distribution networks, a time series is a collection of data points over time, and can provide insights into the behavior and performance of a system. This type of data typically includes variables such as flow rates, pressure levels, water quality parameters, and other relevant metrics recorded at regular intervals [16]. Numerous studies have explored the use of time series data for leak detection in water distribution systems, with examples including modified fuzzy evolving algorithms [14,17]and methods based on parameter determination and performance evaluations [14,18]. However, none of these studies have used a GA to optimize the placement of leakage detection sensors. The analysis of time series data can identify patterns, trends, and anomalies in the behavior of a WDN, and can help to detect leaks, predict water demand, optimize operational performance, and improve the system's overall efficiency.

EPANET

EPANET is a globally used software tool designed to model and simulate water distribution systems [20]. It was primarily developed to enable an understanding of the movement and fate of drinking water constituents within distribution systems; however, its versatility means that it can be applied to various distribution system analysis tasks.

Genetic algorithm

A GA is a non-procedural search technique used in computer science for optimization and search problems to identify the best solution among many potential candidates. This technique was inspired by natural selection and evolution theory, and includes inheritance, mutation, natural selection, and recombination (crossover) operations. Numerous researchers have made significant contributions to WDN leakage detection and localization schemes and have published reviews and shared insights with regard to solutions and recommendations for addressing the challenges associated with WDN leakage [9]. Since 1995, the GA has been utilized as an optimization technique to reduce water leakage in distribution systems [9], and has been used in various applications to detect leaks in WDNs [3,20–22]. Some terms and explanations of the details of the GA are given below [9]:

- Initial population: To initiate the algorithm, a population of solutions is generated. Each potential solution consists of a combination of the number of pipes and the sensors to be installed. A solution is encoded as a binary array, where the length is equal to the number of pipes and each index corresponds to a specific pipe. Creating a solution involves determining the number of sensors to be installed. Indices are then randomly selected, corresponding to the pipes where these sensors will be placed. Each 'gene' (array element) is assigned a value of zero or one in the resulting binary array, where a value of zero indicates that no sensor will be installed on that particular pipe, and a value of one means a sensor will be installed.
- Fitness function: The objective function of the GA is $F(x) = Y = x_1 \times w_1 + x_2 \times w_2 + \ldots + x_n \times w_n$, where x are the gene values, and w are the weights of the genetic algorithm. Each proposed solution will be assessed using two components, i.e. the magnitude and the occurrence of leaks. The main priority is to focus on the occurrence of leaks, with the magnitude of the detected leaks being a secondary concern.
- Selection: To perform mating (crossover and mutation), two solutions are selected (or N solutions are selected in pairs) by choosing the solutions with the highest fitness values.
- Crossover: Candidate binary genes with values of zero and one are selected randomly in one solution, where the selected indices are also random. The crossover operation is applied to the candidate genes using uniform random sampling.
- Mutation: In this process, binary genes with values of zero and one are exchanged randomly within the same solution by selecting indices randomly. This type of operation is referred to as a mutation swap.

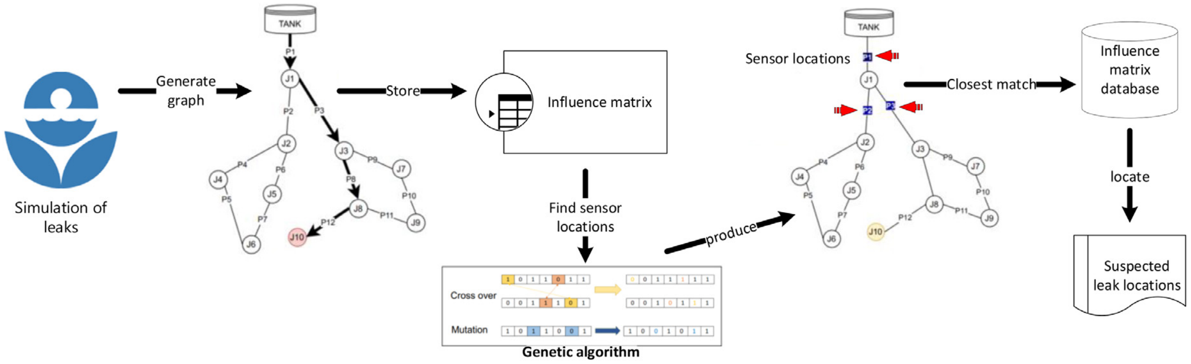

Fig. 1. Architecture of the GA-Sense system.

- Regeneration: Solutions with the lowest fitness are replaced with new solutions with higher fitness scores, to form a new generation of the population.

State of the art

Water leakage is an important issue with far-reaching financial impacts, and the implementation of effective methodologies for leak detection and localization is crucial for preserving water and economic resources. Time series data analysis of WDNs can reveal usage patterns, enabling anomalies to be detected and potential leaks to be identified. The use of a GA for optimal sensor placement can also significantly increase the efficiency of the WDN monitoring process. This proposed GA-based sensor placement strategy, which utilizes time series data, aims to optimize sensor coverage and enhance the effectiveness of leak detection.

Proposed method

Changes in the flow of individual pipes are observed in order to identify leaks in a WDN, and particularly in pipes connected to the reservoir. Current flow centrality is effective for identifying larger leaks, but it is less effective for smaller ones, as these generate less noticeable signatures [23]. Hence, to accurately locate small leaks, it is necessary to analyze more detailed data, such as the flow changes in the surrounding pipes affected by the leak. The proposed method (Fig. 1) simulates leaks and records the changes in pipe flow using an influence matrix. Each solution in the GA is represented as a binary array (gene), where the array length is equal to the number of pipes, and each index corresponds to the position of a pipe. In this array, a gene value of one indicates where a sensor is installed, and zero means that a sensor is not installed. The GA optimizes the sensor locations by employing weighted genes for mutation and crossover tasks. Once the algorithm has identified the optimal sensor locations, leak localization is performed by comparing the observed values from the sensors against a flow database of simulated leaks. The effectiveness of this localization method is evaluated by comparing the suspected leaks with the actual ones; a high rate of matching between these two sets of data indicates the success of the GA-Sense method in terms of accurately detecting and localizing leaks within the WDN. The details of the procedures are explained in the following subsections.

Leak simulation

Many researchers have used the EPANET simulator to develop techniques for detecting leaks, such as in the study in [24]. Leaks are simulated using EPANET by adjusting the emitter coefficients at the network nodes. Initially, these coefficients are set to zero, and to represent a leak with a specific size, the emitter coefficient of a node is modified accordingly. However, accurate simulation of a precise leak size across all nodes can be a complex task, due to the nonlinear dynamics of EPANET, meaning that varying results can be obtained for identical leaks in different pipe orifices. The sensitivity to the emitter coefficient varies with the proximity of a node to the reservoir: nodes closer to the reservoir (in the upstream area) respond more significantly to changes in the emitter coefficient than those further away (in the downstream area). Consequently, a slight increase in the emitter coefficient for upstream nodes can lead to larger leaks than the same adjustment at downstream nodes, due to the higher head pressures in the upstream areas. A trial-and-error approach to finding a proper emitter coefficient can be used based on the hill climbing method. An outline of the hill climbing algorithm and the calibration procedure for the leaks are given below:

A. Hill climbing algorithm: This local search algorithm continuously moves towards an increasing elevation/value to find the peak of a mountain, or in other words the best solution to the problem [25]. It terminates when it reaches a peak value at which none of the neighboring points have higher values. However, the hill climbing algorithm has limitations, as it can become stuck at a local maximum. This situation occurs when the climber reaches a higher peak than all neighboring points but not the highest peak in the mountain range; in this case, the climber must backtrack and explore a different path to reach the highest peak.

Table 2
An illustration of influence matrix.

	P1	P2	P3	P4	P5	P6
j_1	0	0	−0.04	0.07	0	0.05
j_2	0.03	0.05	0.05	0	0	0.05
j_3	0.03	−0.03	0	0	0.03	0
j_4	0.03	0.05	0.02	−0.05	0.03	0
j_5	0.05	0.03	0.05	−0.4	−0.01	0

B. Calibrating leaks using the hill climbing algorithm: To optimize the calibration of leak sizes in the EPANET model, the algorithm works by iteratively adjusting the leak sizes in the EPANET model until the simulated values match the desired values [26]. The hill climbing algorithm in EPANET continuously adjusts the sizes of leaks in a WDN, and compares the simulated results at each iteration with the observed values. Beginning with random leak sizes, the algorithm fine-tunes these estimates and repeats the process until the simulated data closely match the real-world observations, thereby ensuring precise leak modeling.

Characterizing leak signatures

Leak detection and localization typically rely on a leak-free network's minimum night flow (MNF) as a reference point for identifying leaks. The MNF is usually measured in the early morning, when water consumption is at its lowest level, and provides a snapshot of the pipe flows in a leak-free condition. We denote the set flows in pipes of a leak-free network during an MNF as $F = f_1 \ldots f_n$, and the set flows in pipes impacted by a leak at location (node) i as $F_{\mathrm{i}} = fi_1 \ldots fi_n$. The changes in pipe flows resulting from a leak can be expressed as $F - F_i$, which is denoted as $(f_1 - f_1) \ldots (f_n - fi_n)$.

To simplify the notation, we use F_i to symbolize the flow changes in pipes caused by a leak at a specific location i. Since leaks may occur at any spot in a WDN, the number of affected pipes is equal to the node count for the network. The flow changes can be represented as a table to express the impact of a leak on pipe flows (Table 2). The column headings in the table represent flow alterations in pipes, while the row labels represent leak locations. The value in a given pipe can be positive, negative, or zero: a positive value implies an increased flow magnitude, while a negative value means decreased or reversed flow compared to the initial state. Greater positive or negative changes indicate a significant impact of a leak, while values of zero indicate no effect.

Genetic algorithm for placing sensors

The primary objective of GA-Sense is to identify the most effective locations for installing n sensors. This method employs evolutionary principles to iteratively explore and optimize the sensor positions to achieve the best coverage or the optimum detection capability within a given area or system. These optimal sensor locations are assessed based on several specific criteria, as follows:

- Maximum coverage: The strategy focuses on optimizing the placement of sensors to ensure maximum coverage across the entire network. This entails minimizing the number of blind spots and guaranteeing thorough monitoring to achieve comprehensive detection and surveillance throughout the system.
- Sensitivity: The strategic placement of sensor locations is highly sensitive in terms of detecting leaks and other anomalies. Positioning sensors in critical areas ensures the prompt detection of potential issues, enhancing the network's overall reliability and responsiveness to irregularities.
- Cost-effectiveness: The strategy should be efficient in terms of cost, and should place the required number of sensors (n) while still achieving the desired level of network monitoring, without unnecessary expenditure.
- Robustness: The chosen strategy must be robust and capable of adapting to changes in network conditions, such as fluctuations in demand or pressure, without compromising its effectiveness. This criterion ensures reliable long-term performance.

Weighted genes

A design for optimal sensor placement involves positioning sensors on pipes to capture extensive flow data effectively. Any changes in the current values of the influence matrix, whether positive or negative, indicate potential leak events in or around a specific pipe. The weights of the genes are determined using data from this influence matrix. At the evaluation stage, the process is initiated by analyzing the case of a single leak. Each change in the current within the pipes is carefully assessed: positive or negative changes contribute one point to the overall score, whereas a change of zero adds no points. The evaluation proceeds until N leak cases have been considered, resulting in a potential maximum score of N points for a single pipe. These scores, corresponding to the pipes in the influence matrix, are then stored in an array and later utilized as weights in the GA.

A challenge is posed by multiple solutions with identical fitness values, and addressing this issue is important, as such situations can result in unclear selection priorities. To overcome this problem, we introduce a new variable representing the size or magnitude of the leak. We record the scalar size of the changes in current values caused by leaks, and the total scalar size is then averaged over all leak events. The outcome of this process is two fitness arrays, one based on leak events and the other on leak magnitudes. This method improves the precision and effectiveness of our sensor placement strategy and ensures that the optimal sensor locations are determined by the overall network conditions and the nature and magnitudes of the leak events.

Let C be the set of selected column indices and $c \in C$. $T_{:,c}$ is the set of flow changes in a pipe for all leak events. The threshold value for the flow change in a pipe that is considered significant is represented as θ, and the weights for the magnitude and leak events are w_m and w_l, respectively. These two components are then weighted and combined to calculate the overall fitness score ($F(C)$), as shown in Eq. (1).

$$f(C) = w_m \times \sum_{c\in C} \sum_{v\in T_{:,C}} |v| + w_l \times \sum_{c\in C} \sum_{v\in T_{:,C}} 1_{|v|} > 0 \tag{1}$$

where:

- $|v|$ is the absolute value of the flow change;
- $1_{|v|} > 0$ is an indicator function that has a value of one when the absolute value of $|v|$ is greater than the threshold θ, and zero otherwise.

After the fitness of each solution in the initial population has been calculated, the solutions are sorted in descending order based on their fitness. The top N solutions, where N is 10 times the number of sensors, are selected for the next generation. This procedure ensures a good mix of high-quality candidates for crossover and mutation. The process is repeated until the fitness values stabilize, indicating convergence. At this stage, the top solution is chosen as the optimal sensor placement.

Leak detection and localization performance evaluation

When the optimal sensor locations have been identified, the data collected by the sensors for an MNF are matched to a simulated leak database (the influence matrix database, as seen in Fig. 1). Each row in the database contains every possible leak location in a network, along with the flow changes triggered by a leak in the location. Rows with the closest match (measured using the Euclidean distance and coefficient correlation) are selected as suspected leaks. In this study, we consider two leak localization mechanisms: the Euclidean distance and the correlation coefficient.

A. Euclidean distance: In this leak localization method, the distance between two points is calculated in Euclidean space. The result represents the length of a line segment connecting these points. The formula is derived from the Cartesian coordinates of the points, based on the Pythagorean theorem, and is often referred to as Pythagorean distance. When dealing with objects (rather than points), the distance is typically defined as the smallest separation between the pairs of points belonging to the two objects (Eq. (2)).

$$ED(P,Q) = \sqrt{\left(q_{1-}\, p_1\right)^2 + \left(q_{2-}\, p_2\right)^2 + \ldots + \left(q_{n-}\, p_n\right)^2} \tag{2}$$

where P and Q represent two points in an n-dimensional space, and $(q_i - p_i)$ represents the difference between the coordinates of the two points in the ith dimension.

B. Correlation coefficient: This leak localization method quantifies the degree of association between two variables. There are several types of correlation coefficient, with Pearson's being the most widely used. Pearson's correlation (often referred to as Pearson's r) is often used in linear regression analysis (Eq. (3)):

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)(y)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{3}$$

where n is the number of data points or pairs, and x and y represent the data points or pairs.

We evaluate the accuracy of leak localization using two distinct scoring methods: subset and success scoring. In subset scoring, a score of $(^{n}/_{m})$ is assigned when m predicted leak locations include the actual leak among n possible locations. Conversely, in success scoring, a score of one is allocated if any of the predicted leak locations match the actual leak. For instance, if a leak at J_1 is predicted at J_1, J_2, and J_3, subset scoring would give a score of $^{1}/_{3}$, whereas success scoring would yield a score of one. This process is repeated for all potential leak sites, from J_1 to J_n. The accumulated scores are tabulated, and the overall accuracy is computed by dividing the total score by the number of leak cases examined.

Ideal number of sensors

This study focuses on determining the ideal quantity of sensors and their detection effectiveness in a network. To this end, we present a methodology for assessing the impact of varying sensor placements on the efficiency of leak detection. The assessment considers the total number of sensors installed along a pipeline, their precise locations, and the effectiveness of their detection capabilities. Subsequently, an extensive analysis is performed to identify potential trends in detection performance as the quantity of sensors increases. These trends could manifest as an improvement or decrease in returns, or remain consistent. The aim of this investigation is to determine whether there are particular sensor quantities at which detection accuracy stabilizes. Identifying this critical threshold is crucial, as it provides an important reference for determining the most effective number of sensors, tailored to the unique attributes of a given pipeline.

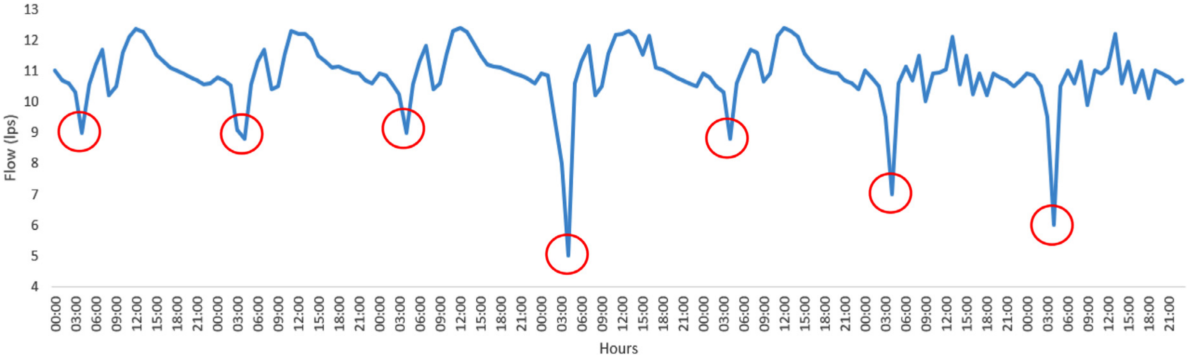

Fig. 2. Fluctuating demand in a water distribution network over a period of one week.

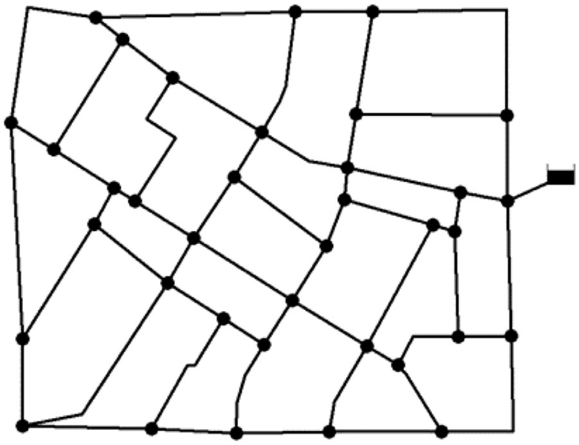

Fig. 3. The Fossiron water distribution network, representing a network with higher degree of connectivity (36 junctions, 58 pipes, and one reservoir, with a connectivity degree of 3.19).

Fluctuation in minimum night flow due to changes in demand

The pattern of flows within the pipes of a network may change over time due to variations in water consumption at demand points. Hence, any analysis of leaks based on flow must consider the changes in this pattern. Several MNF values can be utilized as a reference for a leak-free network when accounting for fluctuating demands. The number of MNFs used during the sensor placement process depends on the magnitude of the fluctuations in the demand in the WDN. If the variations of MNFs is too high, MNFs with the closest values are selected. Large variations in MNFs may occur due to rare events that happen only once within a given timeframe.

We consider the MNFs for a WDN over a period of one week to illustrate the selection process (Fig. 2). The recorded MNF values for each day of the week were as follows: Monday (F_1): 9 lps, Tuesday (F_2): 8.8 lps, Wednesday (F_3): 9.2 lps, Thursday (F_4): 5 lps, Friday (F_5): 8.9 lps, Saturday (F_6): 7 lps, and Sunday (F_7): 6 lps. We note that most MNFs fall within the [8.5, 9.5] lps range, i.e., F_1, F_2, F_3, and F_5. However, F_1 and F_7 are outliers, suggesting that they are potential anomalies due to specific events or other factors. Consequently, F_1, F_2, F_3, and F_5 are adopted as representations of the leak-free state of the WDN. An influence matrix is generated for each snapshot, resulting in a total number of influence matrices that is equal to the number of nodes in the WDN multiplied by the number of snapshots. The sensor placement strategy is then applied using each snapshot.

Experimental results and analysis

Experimental data

Experiments were conducted using the EPANET simulator [19]. The experiments were conducted sequentially, with the proposed sensor placement strategy implemented first and then the leak localization method. Two distinct networks were selected, namely the Fossiron [27] network (Fig. 3) and the Campus XYZ network (Fig. 4), based on their unique characteristics. The Fossiron network was chosen due to its high degree of connectivity, as it features 36 junctions, 58 pipes, and one reservoir, with a connectivity degree of 3.19. In contrast, the Campus XYZ network was selected for its lower connectivity and more dispersed network structure, as it

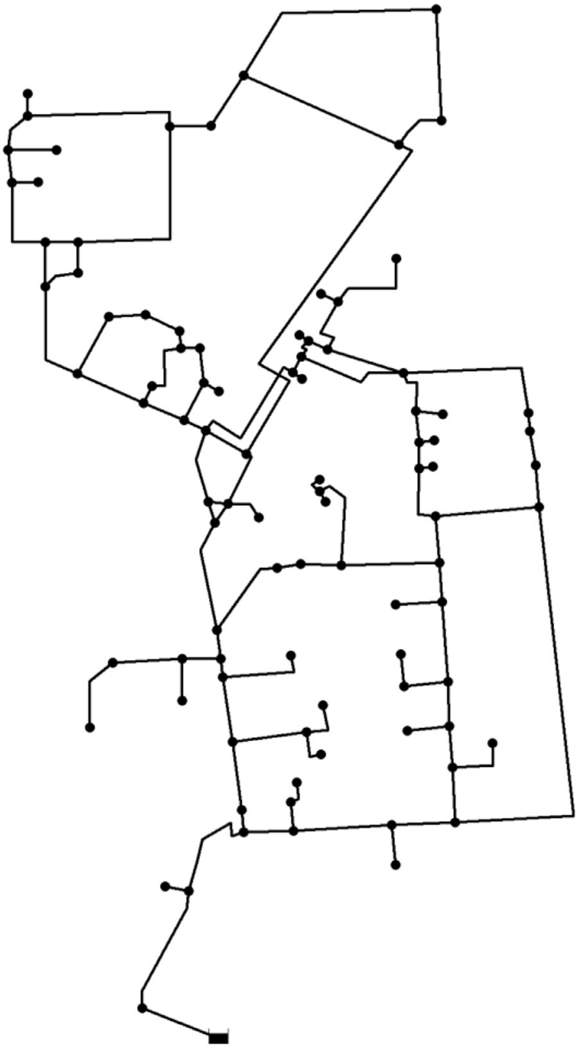

Fig. 4. The Campus XYZ water distribution network, representing a network with lower degree of connectivity and more dispersed network structure (93 junctions, 105 pipes, and one reservoir, with a connectivity degree of 2.5).

includes 93 junctions, 105 pipes, and one reservoir, with a connectivity degree of 2.5. These network choices were made in order to examine the robustness of the proposed GA-Sense in scenarios with varying levels of connectivity and network configurations.

Sensor placement

Genes were chosen for each generation's crossover and mutation processes based on the values of the fitness function. The fitness function was designed with the aim of detecting the maximum number of leaks using as few sensors as possible, while maintaining the optimal detection capability, as determined by the magnitude of delta flow. The results (Table 3) reveal that there may be variations in sensor locations across multiple runs of an experiment and with different sensor quantities. This variability arises from the reliance of the GA on population-based selection in each generation. However, certain locations consistently stand out among the selected set of sensors, indicating their robust leak-detection capabilities. This pattern suggests that these consistently chosen locations offer the strongest detection potential when using a specific number of sensors.

Finding the ideal number of sensors

The leak localization accuracy for each case was measured by comparing the number of correct predictions with the total number of predictions. Two scoring mechanisms were used, subset and success scoring, as explained in Section 3.5. In each experiment, the average localization accuracy was measured by calculating the average localization accuracy levels for all cases of leaks.

Table 3
Sensor locations obtained using GA.

Network	No. of sensors	Generation	Sensor locations	Fitness score
Fossiron	2	7	P14, P58	71
		7	P58, P14	71, 092
	3	7	P14, P54, P58	105
			P14, P52, P58	105
			P14, P53, P58	105
			P14, P15, P58	105
		8	P58, P14, P15	105, 1.167
			P58, P14, P54	105, 1.096
			P58, P14, P53	105, 1.039
			P58, P14, P52	105, 0.985
	4	7	P14, P52, P53, P58	139
			P14, P53, P54, P58	139
			P14, P15, P54, P58	139
			P14, P52, P54, P58	139
			P14, P15, P52, P58	139
		7	P58, P14, P15, P54	139, 1.342
			P58, P14, P15, P53	139, 1.285
			P58, P14, P15, P52	139, 1.230
			P58, P14, P54, P53	139, 1.214
			P58, P14, P54, P52	139, 1.159
			P58, P14, P53, P52	139, 1.103
Campus XYZ	2	7	p13, p62	185
		7	p13, p62	185, 1055
	3	8	p13, p62, p63	275
		8	p13, p63, p62	275, 1.581
	4	7	p13, p28, p62, p63	364
			p13, p39, p62, p63	364
			p13, p62, p63, p112	364
			p10, p13, p62, p63	364
			p13, p30, p62, p63	364
		12	p13, p63, p62, p18	364, 1.847
			p13, p63, p62, p9	364, 1.824
			p13, p63, p62, p36	364, 1.806
			p13, p63, p62, p81	364, 1.755
			p13, p63, p62, p91	364, 1.704
			p13, p63, p62, p52	364, 1.688
			p13, p63, p62, p23	364, 1.673
			p13, p63, p62, p106	364, 1.664

Table 4
Ideal numbers of sensors based on localization performance using the Euclidean distance.

No. of Sensors	Fossiron		Campus XYZ	
	Subset scoring (%)	Success scoring (%)	Subset scoring (%)	Success scoring (%)
2	59.4	100	7.3	100
3	91.6	100	16.3	100
4	91.6	100	22.0	100

Table 5
Ideal numbers of sensors based on localization performance using the coefficient correlation.

No. of Sensors	Fossiron		Campus XYZ	
	Subset scoring (%)	Success scoring (%)	Subset scoring (%)	Success scoring (%)
2	59.4	100	2.1	100
3	91.6	100	3.2	100
4	91.6	100	4.3	100

We varied the number of sensors (two, three, and four) to assess the impact on the performance. Initially, two sensors served as the baseline for evaluating the localization performance, and the performance enhancement achieved by incorporating additional sensors was then investigated. The outcomes consistently revealed improved leak localization accuracy (Tables 4 and 5), and we observed that this performance gap was more noticeable for the Fossiron network. This phenomenon can be attributed to the fact that Fossiron features a higher number of flow pathways (i.e., a higher degree of connectivity), resulting in a more extensive network configuration and, consequently, more significant flow variations.

Table 6
Localization accuracy for fluctuations in MNF.

Network	No. of leak events	Magnitude of leak (l/s)	MNF fluctuations	Average Euclidean distance		Average correlation coefficient	
				Subset scoring (%)	Success scoring (%)	Subset scoring (%)	Success scoring (%)
Fossiron	36	0.5	5%	59.81%	100%	7.39%	100%
			−5%	58.51%	100%	7.39%	100%
			10%	60.74%	100%	7.39%	100%
			−10%	58.51%	100%	7.39%	100%
		1.0	5%	91.67%	100%	17.96%	100%
			−5%	91.67%	100%	17.96%	100%
			10%	91.67%	100%	17.96%	100%
			−10%	91.67%	100%	17.96%	100%
		1.5	5%	91.67%	100%	23.39%	100%
			−5%	91.67%	100%	23.39%	100%
			10%	94.44%	100%	23.39%	100%
			−10%	91.67%	100%	23.39%	100%
Campus XYZ	93	0.5	5%	16.13%	100%	2.15%	100%
			−5%	16.13%	100%	2.15%	100%
			10%	16.13%	100%	2.15%	100%
			−10%	16.13%	100%	2.15%	100%
		1.0	5%	18.28%	100%	3.23%	100%
			−5%	18.28%	100%	3.23%	100%
			10%	18.28%	100%	3.23%	100%
			−10%	18.28%	100%	3.23%	100%
		1.5	5%	37.63%	100%	4.89%	100%
			−5%	37.63%	100%	4.89%	100%
			10%	37.63%	100%	4.89%	100%
			−10%	37.63%	100%	4.89%	100%

At some point, adding sensors to a network does not significantly improve the accuracy of leak detection and localization (as can be seen for the Fossiron network). This suggests that GA-Sense has identified the optimal locations for sensors. Although the leak detection and localization for the Campus XYZ network is low using the subset scoring measurement, the success score reached 100%. This phenomenon also occurs in the Fossiron network, which indicates that suspected leak locations can be narrowed down to a particular area.

Managing fluctuations in minimum night flow

Fluctuations in demand can alter the MNF, consequently impacting the baseline for the proposed sensor placement strategy. The median MNF represents the central point among all values of the MNF. Values were typically observed to fall within a range of up to 10% above or below the median MNF. Based on this assumption, we design experiments with 10% margin above and below this median MNF.

Based on the initial demand pattern, a network was designed that included seven time steps, representing the number of days in a week. To anticipate the demand patterns, we assumed a weekday surge with a 1.5-fold increase in demand from Monday to Friday. This assumption stemmed from the common observation that water demand is typically higher on weekdays than on weekends. The MNF served as the foundational reference value against which deviations in flow patterns, such as leaks, could be identified. The MNF may shift when demand fluctuations occur, meaning that adjustments to the leak detection threshold are needed to accommodate these variations.

From the experimental results (Table 6), we note that the Euclidean distance method produced a better average accuracy than the correlation coefficient method. The superior performance of the Euclidean distance method in finding the most similar vector is due to its ability to consider data points in a multi-dimensional space that accounts for both the magnitude and direction of the differences between data points. This mechanism makes it well-suited to scenarios where similarity is not strictly linear or where complex nonlinear patterns must be detected. In contrast, the correlation coefficient method focuses on linear relationships between two vectors, and is less concerned with the magnitude and direction of differences and more focused on how well the two vectors are linearly related. Since a WDN is a nonlinear system, the Euclidean distance is more suitable for the localization of leaks than the correlation coefficient.

Performance comparison

We evaluated the effectiveness of our proposed GA-Sense method against two other methodologies: leak detection using artificial neural networks (ANNs) [28], and leak detection utilizing the lean graph approach [7,26]. We considered a specific experimental scenario (Table 7) to provide an in-depth evaluation of the accuracy of GA-Sense in comparison to these widely used techniques. In this comparative analysis, we used the Fossiron network, due to its complex structure, as despite being a relatively small network, it

Table 7
Experimental scenario used to compare GA-Sense with other methodologies.

Parameter	Value
Number of sensors	5
Leak size	0.5 lps
Scoring method	Subset scoring
Network	Fossiron
Localization method	Euclidean distance

exhibits a notably high degree of connectivity. We calculated the placement of sensors using each method (GA-Sense, ANN and the lean graph), and performed the localization of leaks using Euclidean distance measurements. This approach represents an experimental scenario similar to our proposed method (see Subsection 3.5). We employed subset and success scoring to enable a rigorous assessment of the leak localization performance for five sensors within the Fossiron network.

The experimental results were excellent for all methods, with values 100% for the success score. This result indicates that all three methods could localize leaks within the suspected locations. When assessed using subset scoring, GA-Sense outperformed the other methods, achieving the highest accuracy of 91.6%, while the lean graph method and ANN produced lower accuracies of 84.5% and 46.5%, respectively. These findings illustrate the superiority of GA-Sense in terms of strategically placing sensors.

Discussion

The GA performs well when used to find sensor placements for leak detection. The effectiveness of the proposed method was proven for various network configurations, and the Euclidean distance formula was found to be superior to the correlation coefficient in detecting and localizing leaks, due to its ability to handle multi-dimensional data variability. In general, the proposed method is adaptable to diverse types of water systems and conditions. Integrating the proposed method with advanced data acquisition using IoT technologies would enable more efficient and effective real-time monitoring and analysis of water distribution networks. However, as the effectiveness of the proposed method primarily depends on the quality and accuracy of the data, challenges may arise due to the potential non-universality of the weekday demand surge in a real-world situation. Exploring the performance of our method using various changing network conditions would be beneficial to enhance its efficacy and reliability.

Conclusion

This paper has presented two significant advancements in WDN management. Firstly, we developed a novel GA fitness function using a weighting scheme to select sensor locations, which considers the frequency and magnitude of leak events in order to avoid the common pitfall of GA settling for local optima, thus ensuring a more globally optimal solution for sensor placement. Secondly, the proposed method can dynamically identify the most reliable sensor locations over time. This enables the sensor network to adapt to changing conditions within the water distribution system, such as fluctuations in water demand or alterations to the network structure. Our experimental results showed that GA-Sense could effectively identify suspected leak locations within a specific area, as it achieved a 100% success score. These contributions will be instrumental in advancing the field of WDNs by providing a robust framework for efficient leak detection and localization. There are promising avenues for further research in this area, such as testing the real-world efficacy of our sensor placement strategy in water networks and exploring its adaptability to demand fluctuations, network changes, and sensor reliability in order to enhance long-term optimization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ary Mazharuddin Shiddiqi: Conceptualization, Methodology, Investigation, Writing – original draft. **Choiru Za'in:** Software, Validation. **Artya Lathifah:** Methodology, Writing – review & editing. **Tohari Ahmad:** Validation, Data curation. **Diana Purwitasari:** Project administration, Resources.

Data availability

Data will be made available on request.

Acknowledgments

Funding: This work was supported by the Annual Funding of the Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

References

[1] O. Bello, A. Abu-Mahfouz, Y. Hamam, P. Page, K. Adedeji, O. Piller, Solving management problems in water distribution networks: a survey of approaches and mathematical models, Water (Basel) 11 (3) (Mar. 2019) 562, doi:10.3390/w11030562.

[2] X. Xie, Q. Zhou, D. Hou, H. Zhang, Compressed sensing based optimal sensor placement for leak localization in water distribution networks, J. Hydroinf. 20 (6) (Nov. 2018) 1286–1295, doi:10.2166/hydro.2017.145.

[3] M.V. Casillas, V. Puig, L.E. Garza-Castañón, A. Rosich, Optimal sensor placement for leak location in water distribution networks using genetic algorithms, Sensors (Basel) 13 (11) (Jan. 2013) 14984–15005, doi:10.3390/s131114984.

[4] L.E. Garza-casta, "Optimal sensor placement for leak location in water distribution networks using genetic algorithms ˜," pp. 14984–15005, 2013, 10.3390/s131114984.

[5] M. Kammoun, A. Kammoun, M. Abid, Leak detection methods in water distribution networks: a comparative survey on artificial intelligence applications, J. Pipeline Syst. Eng. Pract. 13 (3) (Aug. 2022), doi:10.1061/(asce)ps.1949-1204.0000646.

[6] Z. Hu, D. Tan, B. Chen, W. Chen, D. Shen, Review of model-based and data-driven approaches for leak detection and location in water distribution systems, Water Supply 21 (7) (Nov. 01, 2021) 3282–3306 IWA Publishing, doi:10.2166/ws.2021.101.

[7] A.M. Shiddiqi, R. Cardell-Oliver, A. Datta, An advanced sensor placement strategy for small leaks quantification using lean graphs, Water (Basel) 12 (12) (Dec. 2020) 3439, doi:10.3390/w12123439.

[8] R. Vanijjirattikhan, et al., AI-based acoustic leak detection in water distribution systems, Result. Eng. 15 (2022), doi:10.1016/j.rineng.2022.100557.

[9] S. Katoch, S.S. Chauhan, V. Kumar, A review on genetic algorithm: past, present, and future, Multimed. Tools Appl. 80 (5) (Feb. 2021) 8091–8126, doi:10.1007/s11042-020-10139-6.

[10] J.S. Arora, Multi-objective optimum design concepts and methods, in: Introduction to Optimum Design, Elsevier, 2012, pp. 657–679, doi:10.1016/B978-0-12-381375-6.00017-6.

[11] M. Casillas, V. Puig, L. Garza-Castañón, A. Rosich, Optimal sensor placement for leak location in water distribution networks using genetic algorithms, Sensors 13 (11) (Nov. 2013) 14984–15005, doi:10.3390/s131114984.

[12] S. Khalifeh, S. Akbarifard, V. Khalifeh, E. Zallaghi, Optimization of water distribution of network systems using the Harris Hawks optimization algorithm (Case study: homashahr city, MethodsX 7 (2020) 100948, doi:10.1016/j.mex.2020.100948.

[13] R. Peirovi Minaee, M. Afsharnia, A. Moghaddam, A.A. Ebrahimi, M. Askarishahi, M. Mokhtari, Calibration of water quality model for distribution networks using genetic algorithm, particle swarm optimization, and hybrid methods, MethodsX 6 (2019) 540–548, doi:10.1016/j.mex.2019.03.008.

[14] W.D. Ray, D. Brillinger, Time series: data analysis and theory, J. R. Stat. Soc. Ser. A. 145 (1) (1982) 134, doi:10.2307/2981427.

[15] Y. Shao, X. Li, T. Zhang, S. Chu, X. Liu, Time-series-based leakage detection using multiple pressure sensors in water distribution systems, Sensors (Switzerland) 19 (14) (2019), doi:10.3390/s19143070.

[16] B. Ferreira, N. Carriço, R. Barreira, T. Dias, D. Covas, Flowrate time series processing in engineering tools for water distribution networks, Water Resour. Res. 58 (6) (Jun. 2022), doi:10.1029/2022WR032393.

[17] L. Birek, D. Petrovic, J. Boylan, Water leakage forecasting: the application of a modified fuzzy evolving algorithm, Appl. Soft Comput. J. 14 (2014) PART B, doi:10.1016/j.asoc.2013.05.021.

[18] T. Zhang, X. Li, S. Chu, Y. Shao, Parameter determination and performance evaluation of time-series-based leakage detection method, Urban Water J 18 (9) (2021), doi:10.1080/1573062X.2021.1930067.

[19] L.A. Rossman, EPANET 2: Users Manual and others, U.S. Environmental Protection Agency, Cincinnati, 2000.

[20] J.P. Vitkovsky, A.R. Simpson, M.F. Lambert, Leak Detection and calibration using transients and genetic algorithms, J. Water Resour. Plan Manag. 126 (4) (2000) 262–265, doi:10.1061/(ASCE)0733-9496(2000)126:4(262).

[21] A.R. Simpson, G.C. Dandy, L.J. Murphy, Genetic algorithms compared to other techniques for pipe optimization, J. Water. Resour. Plan Manag. 120 (4) (1994) 423–443, doi:10.1061/(ASCE)0733-9496(1994)120:4(423).

[22] Z.Y. Wu, P. Sage, Water loss detection via genetic algorithm optimization-based model calibration, Water Distribut. Syst. Anal. Sympo. 247 (2005) (2006) 180 *2006*, doi:10.1061/40941(247)180.

[23] L. Romero-Ben, D. Alves, J. Blesa, G. Cembrano, V. Puig, E. Duviella, Leak detection and localization in water distribution networks: review and perspective, Annu. Rev. Control 55 (2023), doi:10.1016/j.arcontrol.2023.03.012.

[24] A. Nagaraj, G.R. Kotamreddy, P. Choudhary, R. Katiyar, B.A. Botre, Leak detection in smart water grids using EPANET and machine learning techniques, IETE J. Educ. 62 (2) (2021), doi:10.1080/09747338.2021.1984317.

[25] A.W. Johnson, S.H. Jacobson, On the convergence of generalized hill climbing algorithms, Discrete Appl. Math. 119 (1–2) (2002) 37–57 *(1979)*Jun., doi:10.1016/S0166-218X(01)00264-5.

[26] A.M. Shiddiqi, R. Cardell-Oliver, A. Datta, Sensor placement strategy for locating leaks using lean graphs, in: *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, in CySWATER2017, Pittsburgh, Pennsylvania, USA, ACM, 2017, doi:10.1145/3055366.3055372.

[27] National Institute for Hometown Security, " Water Distribution Systems Toolkit.", 2016, Date accessed: 25/01/2024, https://www.uky.edu/WDST

[28] X. Fan, X. Zhang, X.B. Yu, Machine learning model and strategy for fast and accurate detection of leaks in water supply network, J. Infrastruct. Preservat. Resilience 2 (1) (2021), doi:10.1186/s43065-021-00021-6.