UNIVERSITÀ DEGLI STUDI DI NAPOLI "PARTHENOPE"

Dipartimento di Scienze e Tecnologie (DiST)
Dipartimento di Ingegneria

PhD Program in

Fenomeni e Rischi Ambientali (FeRiA)

XXXVIII Cycle

Thesis Title

Mitigation of Hydrogeological Risk Caused by Leakage in Urban Water Distribution Networks: *Sensor Placement Optimization Methods*

Author:
Gabriele Medio
PhD Candidate

Tutor:
Prof. Renata Della Morte

Co-Tutor:
Prof. Luca Cozzolino
Dr. Giada Varra

PhD Program Coordinator:
Prof. Antonio Occhiuzzi

ACADEMIC YEAR 2024/2025

Quando soffia il vento del cambiamento,

alcuni costruiscono muri,

altri mulini a vento

Table of contents

Abstract

Water losses in urban water distribution networks not only represent a waste of a fundamental resource but also generate a further, often overlooked, negative impact: they act as territorial stress factors capable of triggering hydrogeological instability phenomena, such as ground settlements and sinkholes. The thesis focuses precisely on this second aspect, highlighting and addressing an anthropogenic risk that is frequently neglected and undervalued, yet proves to be a real and increasingly concerning threat to public safety, transport networks, and structural integrity, standing as one of the most subtle and insidious challenges of the modern urban context.

To address this challenge, the study presents an innovative, integrated multidisciplinary framework designed to mitigate the Hydrogeological Disruption caused by Leaks (HDL) risk in urban water distribution networks. The proposed approach interconnects within a single operational framework components usually treated independently: territorial risk zoning, water leak localization, and pressure sensor placement optimization. Thanks to its modular nature, the methodology allows for the replacement of individual steps with higher-performing techniques or future evolutions based on specific problem requirements. The goal is to maximize leak localization accuracy in the most exposed and critical urban areas, providing the foundation for more targeted interventions aimed at mitigating the risk in question.

The framework was tested on the *Real Network 1* and *L-Town* case studies. The results demonstrate that integrating the HDL risk component promotes a more balanced and effective sensor distribution for urban territory protection. Such a configuration does not require significant deviations from classic hydraulic schemes, ensuring economic sustainability and ease of implementation for water utilities. In conclusion, this work proposes a flexible and scalable model to innovate water network management through a proactive approach toward urban hydrogeological disruptions, in line with modern sustainability paradigms. The added value lies in considering the network not as an isolated system, but as a significant factor for urban resilience, highlighting the hitherto neglected necessity of monitoring the interaction between hidden leaks and the stability of the urban fabric.

1. Introduction

Managing water resources effectively is becoming one of the most significant challenges of the 21st century. Around the world, water supply systems are facing increasing pressure due to a combination of factors including population growth, climate change, rapid urban development, and the progressive deterioration of infrastructure.

Urban water distribution networks play a key role in ensuring a continuous, safe and reliable supply of drinking water. However, these systems are often affected by ageing components, insufficient maintenance and widespread structural weaknesses, which make them particularly prone to hidden leaks. These losses not only result in a substantial waste of water, but also contribute to a critical and often underestimated issue: hydrogeological instability in urban environments, which can manifest as ground subsidence, surface deformation or the formation of sinkholes.

Although the relationship between water leaks and soil instability has not been extensively investigated in the scientific literature, it is of considerable importance. Such processes can cause serious damage to infrastructure, disrupt traffic circulation, damage vehicles, compromise underground utilities, weaken the structural integrity of buildings and, in the most severe cases, pose a threat to public safety.

For these reasons, the management of modern water networks should not be limited to improving hydraulic efficiency. It should also incorporate strategies aimed at reducing hydrogeological risks associated with pipeline leaks, within a broader vision of sustainable development, urban safety and responsible territorial management.

This study presents an innovative multidisciplinary framework for mitigating the hydrogeological risks caused by water leakages. The proposed approach combines the optimisation of pressure sensor placement with a spatial risk zonation methodology. The objective is to maximise leak localisation accuracy in the most vulnerable and exposed areas of the city, supporting proactive interventions and enhancing the overall resilience of the urban system.

1.1. Problem Relevance

Water loss represents one of the most persistent and concerning inefficiencies affecting urban distribution systems. The problem is widespread and affects a broad range of infrastructures and territories, often very different in terms of development and management. In many cases, the volumes of water lost are far from negligible.

On a global scale, it is estimated that roughly 126 billion cubic meters of water are lost every year through urban networks, with an associated economic impact exceeding 39 billion US dollars. [1] In developing countries, water losses through leaks alone are estimated at 45 million cubic meters per day, a volume sufficient to supply nearly 200 million people. In addition, an estimated 30 million cubic meters per day are not invoiced due to unauthorized consumption, corruption, or metering errors. [2] An overview of this critical issue is presented in Figure 1, which illustrates the percentage of Non-Revenue Water (NRW) by country using a color-coded scale to highlight the most affected areas worldwide.

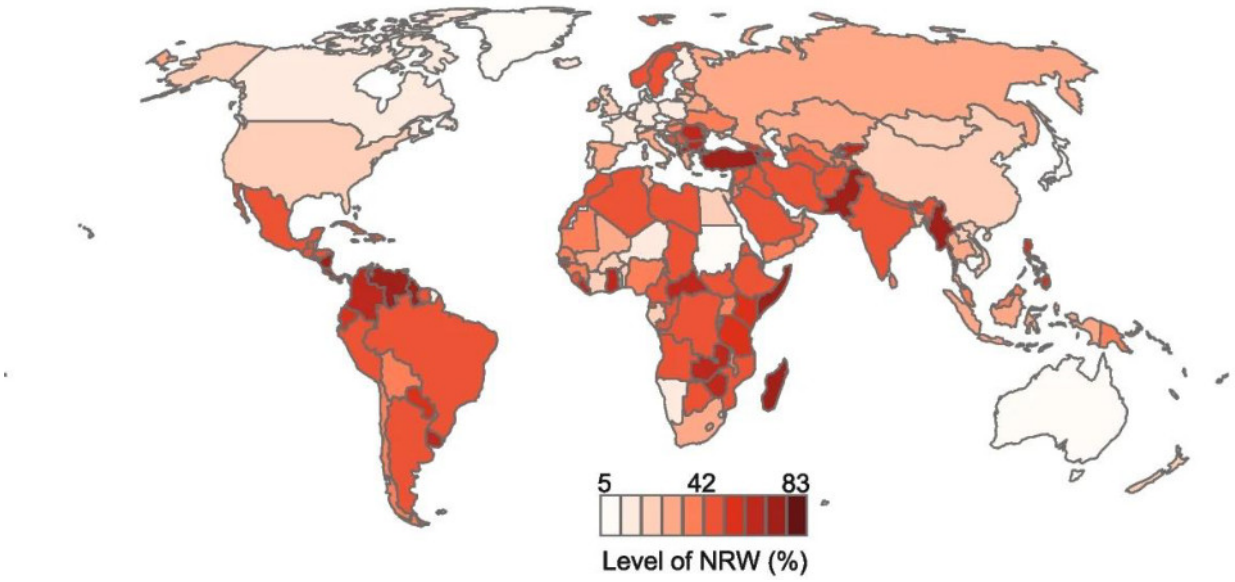

Figure 1. Reproduced from Evaristo et al. (2023), Sustainable Earth Reviews. [3] Global distribution of Non-Revenue Water (NRW) levels by country, shown as a percentage of the total system input volume. The map uses a color gradient to indicate severity: darker shades correspond to higher water loss rates, highlighting areas with the most critical inefficiencies.

Across Europe, public water networks lose on average around 26 percent of the water they distribute. However, this figure is not uniform, and some countries show significantly higher or lower values. [4] A more detailed distribution of these losses is illustrated in Figure 2.

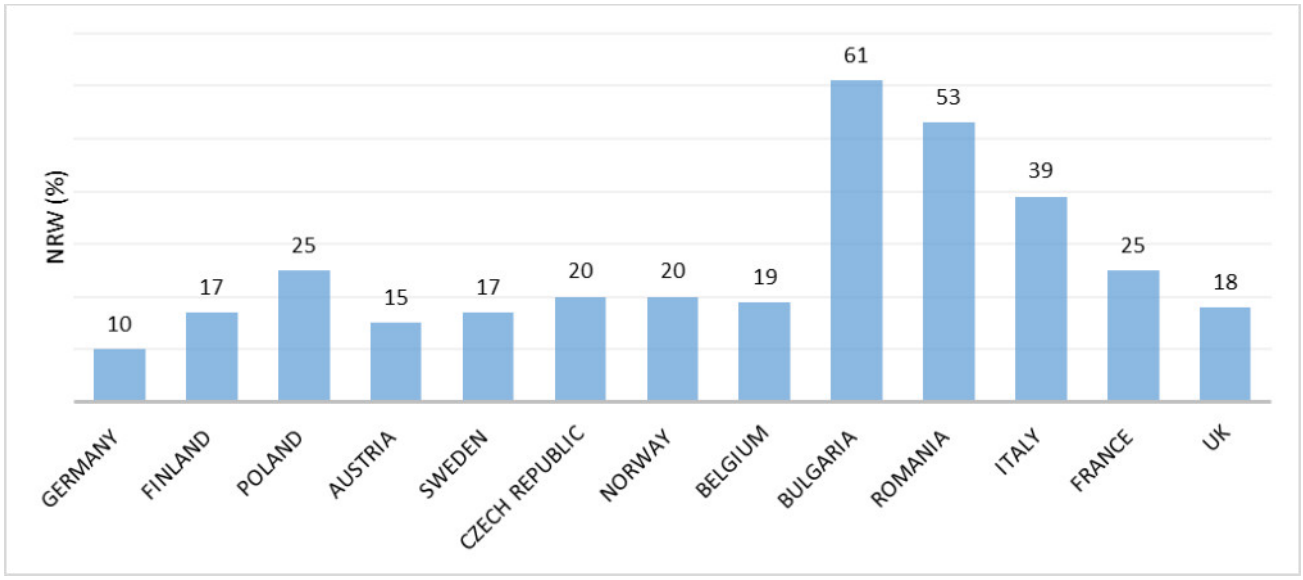

Figure 2. Estimated percentages of Non-Revenue Water (NRW) across selected European countries. The chart highlights significant disparities, with Bulgaria and Romania showing the highest rates (61% and 53%, respectively), while countries like Germany (10%) and Austria (15%) maintain much lower levels. Adapted from AVK Group, AVK Non-Revenue Water Solutions brochure, 2017 [4].

Shifting the focus from the global and European scale to the Italian scenario, the situation can be described by the following data. In 2020, according to ISTAT (Italian National Statistics Institute) data, water losses in the distribution phase reached 3.4 billion cubic meters, equal to 42.2% of the water fed into the network, equivalent to the annual needs of over 43 million people. Compared to 2018, the value is substantially stable (42.0%), confirming the persistent inefficiency of municipal networks. [5]

The losses show a marked territorial disparity, with higher levels in the regions of Central and Southern Italy. The highest values are recorded in Basilicata (62.1%), Abruzzo (59.8%), Sicily (52.5%) and Sardinia (51.3%), while the lowest belongs to Valle d'Aosta (23.9%). In the river basin districts of Sicily (52.5%), Sardinia (51.3%), the Southern Apennines (48.7%) and the Central Apennines (47.3%), total water losses in distribution reach the highest levels at national level. On the other hand, in the northern districts there are lower values, in particular in that of the Po River, the indicator records the lowest value, equal to 31.8% of the volume injected into the network. [5]

At the municipal level, more than half of Italian municipalities (57.3%) have losses of more than 35%, and in one in four dispersions exceed 55%. In the provincial capitals, the average drops to 36.2%, but more than half of the regions show an increase in losses compared to 2018, in particular Basilicata, Molise and Abruzzo. [5]

The consequences of such inefficiencies go well beyond economic loss. They also contribute to unnecessary energy usage, mismanagement of available water, and a general decline in the reliability of water services, particularly during periods of drought or in emergency contexts. [6]

It is estimated that in 2050 the global urban population experiencing water scarcity will increase from one-third to almost half of the global urban population, or from 933 million (referring to 2016) to 1.693-2.373 billion people. [7] Therefore, the problem of water loss carries implications that extend far beyond system performance. Addressing these inefficiencies is also essential to ensure that urban areas can remain functional and adaptable in the face of increasing environmental pressures and population growth.

Unfortunately, the effects of water losses do not end with the waste of resources, which in itself is already a major problem. The leakage of water from underground pipes can alter the geotechnical conditions of the soil, triggering instability phenomena in the ground and in the urban fabric, with consequent deformations, differential subsidence or the formation of sinkholes (see Figure 3). These processes can cause damage to infrastructure, traffic disruptions, damage to vehicles and compromise of underground networks, as well as weakening the structural integrity of buildings. In severe cases, such events can pose a serious threat to public safety.

The phenomena of hydrogeological instability induced by water losses from underground pipelines in urban environments are not yet widely covered in scientific literature, although several studies have analyzed the dynamics of these processes in depth. However, with the increasing number of subsidence recorded in urban contexts, the study of the mechanisms of ground collapse associated with leaks in groundwater networks is becoming increasingly important [8].

The causal connection between leaks and instability has been recognized in several geotechnical studies and laboratory experiments, which show how the erosion of the soil surrounding the pipe can trigger localized subsidence and the formation of underground cavities. Some of these studies are reported below.

D'Aniello et al. (2021) analyze the interaction between water losses and subsurface dynamics, highlighting how these phenomena can modify water paths and promote urban karst processes within underground utility trenches. The study shows that a significant share of the dispersed volume (between 55% and 73%) tends to remain trapped in the trench itself, without reaching the underlying aquifer in about a third of the simulated cases. [9]

Ali and Choi (2021) developed a series of small-scale physical models to analyze the formation of anthropogenic sinkholes caused by leaks in underground pipelines. Experiments evaluated the effect

of different subsurface stratigraphic profiles, flow conditions and leak location on the speed and magnitude of ground subsidence. The results showed that soil stratigraphy is the dominant factor in the mechanisms of sinkhole formation, while other parameters contribute secondarily to the subsidence phenomenon. In particular, the profiles consisting of sandy clay, limestone and bedrock (SC–LS–BR) were found to be the most vulnerable to the development of cavities. The experimental data collected were used for the definition of a sinkhole risk index (SRI) useful for the assessment and mapping of potentially unstable areas. [10]

As can be seen below, many studies on hydrogeological instability caused by water leaks in underground water pipelines have been conducted in China, it is no coincidence that in recent years a significant increase in cases of urban ground collapse (UGC) has been observed on its territory, with serious economic and social consequences. The phenomena covered by this study, however, can also be observed in other countries, although the attention to these problems unfortunately appears to be less. The analysis conducted by Wang and Xu (2022) shows that the frequency of these events is higher in eastern coastal areas, characterized by high urbanization, and lower in northeastern regions. Factors contributing to the collapse include both natural causes, such as geological conditions and heavy rainfall, and anthropogenic causes, including groundwater pumping, leaks from underground pipelines and underground construction. In particular, changes in the groundwater regime play a central role in modifying the stability of soils, especially in the presence of silty soils or karst areas. The authors emphasize the importance of integrated groundwater management and preventive measures such as detailed geological surveys, sponge city design, and multi-service tunneling construction to mitigate the risk of urban collapse. [11]

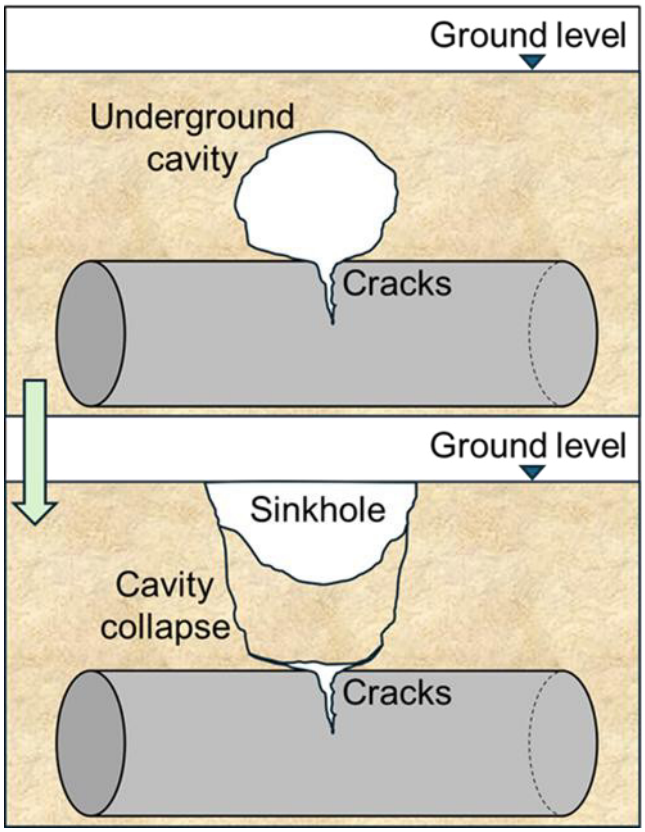

Figure 3. Conceptual scheme of the process of formation and collapse of an underground cavity induced by leaks or defects in an underground pipeline. The infiltration of water and soil particles through the cracks in the pipeline leads to the progressive formation of a cavity in the overlying soil. As the volume of the cavity increases and the bearing capacity of the soil is reduced, the collapse of the ceiling is triggered with the consequent formation of a sinkhole on the surface. Own elaboration inspired by Karoui et al. (2018) [12].

In particular, Mao et al. (2023) investigated the erosive effects of leaks from the water supply network on clayey-type road foundations, identifying water flow rate, leak opening size and emission angle as the main parameters. Experimental results showed that a reduction in discharge contributes significantly to increasing soil erosion resistance, while also providing a detailed description of the progressive evolution of the cavity in the soil. [13]

Dastpak et al. (2023) offer an in-depth review of the phenomenon of Soil Erosion due to Defective Pipes (SEDP), highlighting that it is one of the main causes of artificial sinkholes and land subsidence in urban areas. The authors point out that, although natural sinkholes are linked to karst geological processes, those that occur in urban contexts derive mainly from water losses due to the deterioration of pipelines. These events, exacerbated by climate change, storm surges and increasing urbanization, pose an increasing risk to infrastructure and densely populated areas. [14]

Moreover, it has been highlighted by Cui et al. (2024) that the water pressure inside the pipeline is a key factor in determining the shape and intensity of instability in the event of leaks. High pressures produce significant soil erosion, with the formation of internal cavities and marked surface subsidence, while lower pressures cause ellipsoidal water diffusion and the appearance of circular fissures on the surface, as can be seen in Figure 4. [15]

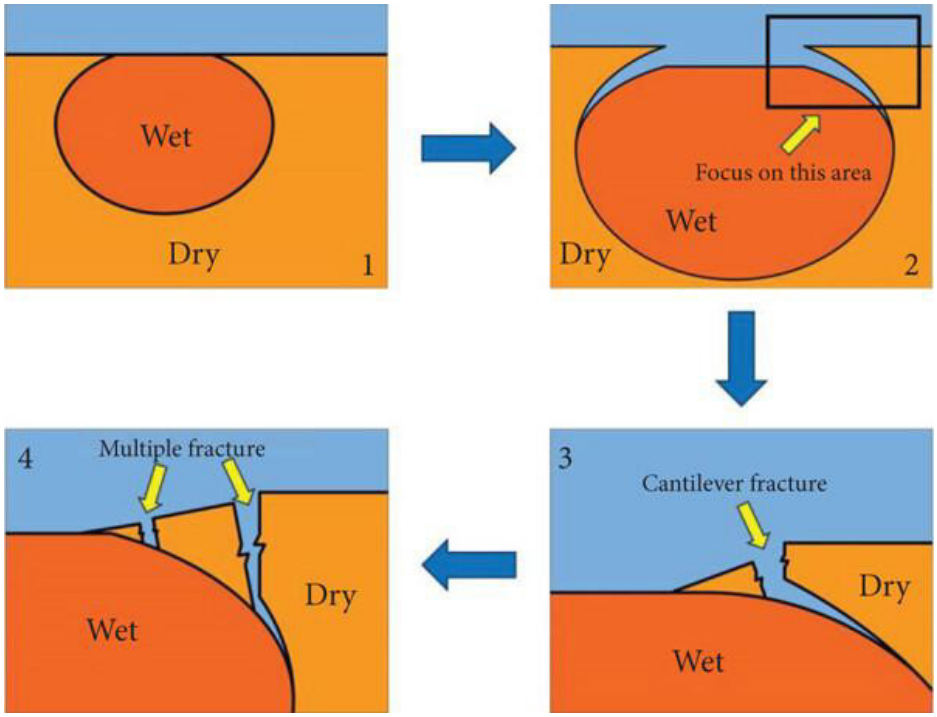

Figure 4. Reproduced from Cui et al. (2024) [15]. Diagram of the process of formation of fractures in the ground due to the ellipsoidal diffusion of water following the leakage from a pressure pipeline. Moisture propagates from the point of damage, generating a saturated (ellipsoidal) zone subject to subsidence; the portions of soil above behave like small rigid corbels which, as the downward movement increases, progressively fracture as the effect of their own weight.

An experimental study by Guo et al. (2024) reproduced the full-scale conditions of a road failure due to underground pipeline leakage in Anqing City, China, using a large-scale three-dimensional test rig (3 × 2 × 2 m). Using Digital Image Correlation (DIC) technology, the authors observed that groundwater percolation is the main driving force of the failure, while defects in the pipeline act as trigger points, creating migration spaces for the eroded soil. When the water flow is reduced, the

cohesive forces between soil particles maintain stability; however, as the infiltration increases, the soil structure progressively weakens, leading to soil failure. [8]

Liu et al. (2024) used an experimental and numerical approach based on the Computational Fluid Dynamics–Discrete Element Method (CFD–DEM) to analyze the soil failure mechanism due to water leakage from shallow water pipes. Through physical model tests, the authors examined the effect of variables such as defect size and location, internal pressure, embedment depth, and groundwater level. The study identified three distinct soil behaviors: mild disturbance, erosional cavity formation and collapse, and soil fluidization. It was found that fluidization occurs when the ratio of hydraulic head to embedment depth exceeds the critical value of 2, while the migration of fine particles around the defect promotes cavity enlargement and instability. [16]

An example of road collapses caused by underground pipe leaks can be seen in Figure 5, while a statistical analysis of the main causes of road collapse accidents (RCAs) is reported in Figure 6. [16]

Figure 5. Reproduced from Liu et al. (2024) [16]. Typical road collapses due to leaking underground pipes in China: left, seepage from a drainage pipe; center, water leaking from a rainwater pipe; right, rupture of a pressurized water pipe (image by Yi-Chun Cao).

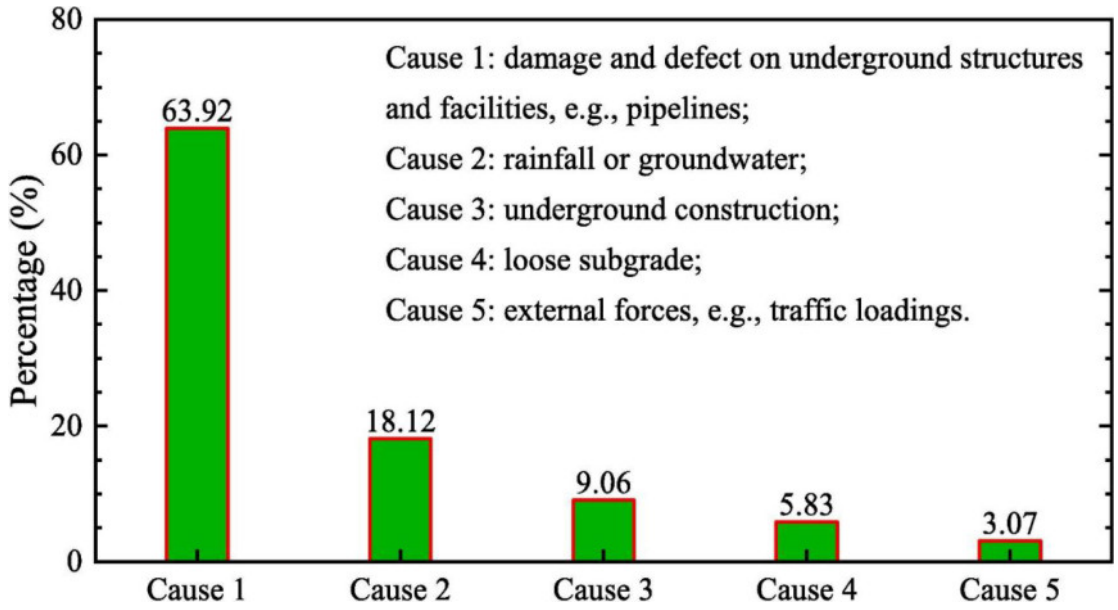

Figure 6. Reproduced from Liu et al. (2024) [16]. Statistical analysis of the main causes of the 620 road collapse accidents (RCAs) recorded between 2017 and 2021. The data shows that damage and defects to underground structures and infrastructures, particularly shallow and defective underground pipelines, are responsible for approximately 63.9% of the analyzed events. The other identified causes include rain or groundwater infiltration, underground construction activities, poor subgrade quality, and external stresses due to traffic.

A similar approach to the previous one was used by Guo et al. (2025). Through large-scale physical tests and simulations (CFD–DEM), the authors analyzed soil–water interactions during seepage erosion. The results show that water infiltration promotes particle movement and that flow velocity is a key factor in exceeding the soil disintegration threshold. The thickness of the cover layer only affects the collapse time, but not the erosion rate. [17]

Instead, Chao et al. (2025) analyzed the ground failure mechanisms associated with the sudden rupture of underground pressurized water pipes in an urban environment (Figure 7). The study, based on laboratory experiments and a coupled solid–fluid Finite Difference Method-Discrete Element Method (FDM-DEM) numerical analysis, highlighted that the failure process develops in three main phases: infiltration diffusion, expansion of the erosion cavity, and soil fluidization. Through Digital Image Correlation (DIC) analysis, the authors identified a wedge-shaped displacement zone, characterized by maximum shear stresses at the edges. It was also shown that the emplacement depth, the location of the leak, and the internal pressure influence the propagation of the cavity and the maximum water escape distance. Finally, microscopic scale analyses showed that fine particles are more vulnerable to erosion, as they form less resistant bonds with surrounding particles. [18]

Internal soil erosion caused by leaks in faulty underground pipelines is a complex process, strongly influenced by the position of the defect along the pipeline. In a combined experimental and numerical study, Wang et al. (2025) analyzed the effects of different leak locations on soil and water loss (SWL) and the resulting soil failure. Experiments showed that by moving the location of the defect from the upper part of the pipeline (crown) to the lower part (invert), the erosion rate and the severity of the failure significantly increase. Simulations, also performed with an FDM–DEM model, highlighted how variations in soil and water pressure around the defect favor local instability. It was also observed that the lateral pressure distribution can be divided into three characteristic zones: fluctuation, arch effect and stability. Here too, the results provide a useful contribution to understanding the mechanisms of internal erosion and soil failure due to localized leaks in underground pipelines. [19]

Figure 7. Reproduced from Chao et al. (2025) [18]. Two typical urban pipe burst accidents: on the left, the event in Shanghai, China (November 11, 2017); on the right, the event in Zhengzhou, China (December 6, 2013).

Studies have also been conducted on the mechanisms of ground failure induced by leaks and defects in sewer pipes, such as those of Karoui et al. (2018), where they analyzed how such anomalies can generate underground voids and, in the most severe cases, surface collapses. Through a physical model of a damaged pipe, the authors identified the direction of groundwater flow, the hydraulic gradient around the leak point, and the soil resistance as determining factors, which influence soil deformation, cavity propagation, and collapse speed. It was also observed that pore pressure varies

significantly during the expansion and collapse phases of the cavity. Although the research refers to sewer pipes, the results can be partially extended to pressurized water pipes, where similar phenomena of internal erosion and ground instability occur. [12]

In an in-depth review, Ali and Choi (2019) analyzed leak monitoring methods and associated sinkholes, with a focus on the use of Wireless Sensor Networks (WSNs). The study compares different approaches, including patent analysis, literature searches and WSN applications, highlighting that wireless sensor-based monitoring technology is still at an early stage of development and requires further experimentation. The authors suggest integrating WSNs with the Internet of Things (IoT) and artificial intelligence in order to create more efficient and predictive systems for early leak detection and sinkhole prevention. [20]

Furthermore, Ali and Choi (2019) provide a conceptual framework of the causes and effects associated with leaks and breaks in underground water pipes, as shown in Figure 8. [20]

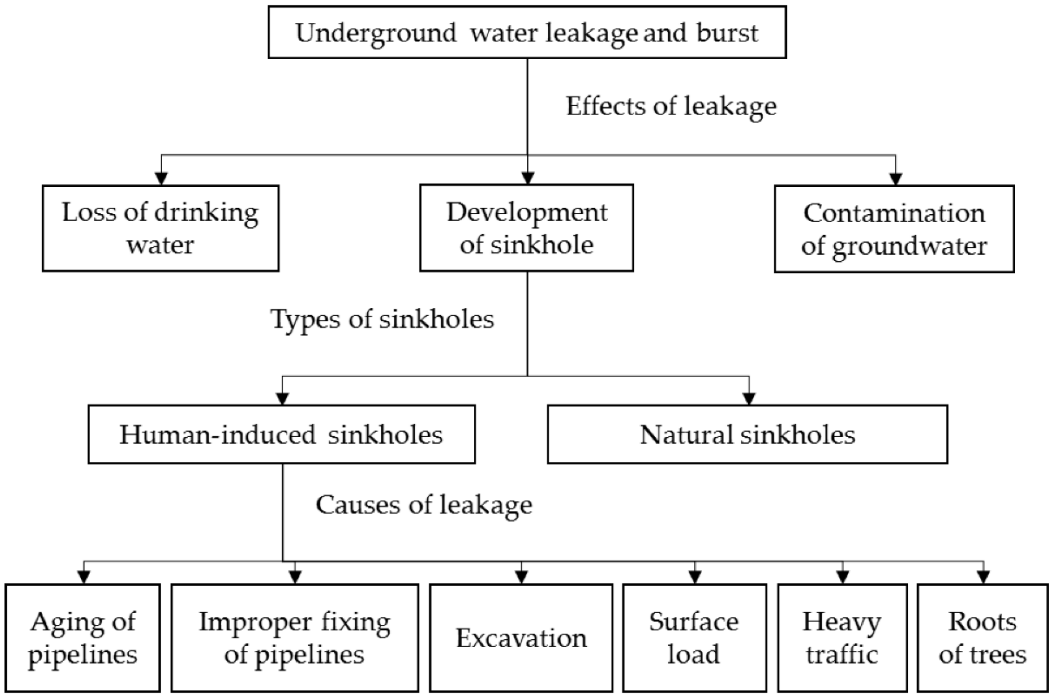

Figure 8. Reproduced from Ali and Choi (2019) [20]. Conceptual framework of the effects and causes associated with leaks and breaks in underground water pipes. Leaks can lead to the dispersion of drinking water, contamination of the groundwater, and the development of sinkholes. Sinks can be divided into natural and anthropogenic, with the latter frequently linked to infrastructural factors such as age or incorrect pipe installation, excavations, surface loads, heavy traffic, or tree roots. In this paper, particular attention is paid to anthropogenic sinkholes, generated by water leaks in urban distribution networks.

In Italy, studies dedicated to anthropogenic sinkholes are still limited; however, a significant contribution is represented by the work of Tufano et al. (2022), which analyses in detail the situation in the city of Naples, frequently affected by ground collapse phenomena (Figures 9 and 10). The study updated the existing inventory by including 270 new events that occurred between 2010 and 2021, for a total of 458 documented cases since 1880. The analysis highlighted a greater concentration in the historic center, where the presence of cavities and underground tunnels favors ground collapse, often in conjunction with intense rainfall or damage to service networks. A direct correlation between

monthly rainfall and sinkhole activation also emerged, and a preliminary susceptibility assessment using the Frequency Ratio method identified the central sector of the city as the most vulnerable. [21]

The sinkhole on Via Campanile, in the Pianura neighborhood of Naples (Figure 11), is a prime example of urban collapse induced by water leaks and infrastructure fragility in Italy. The event, which occurred in February 2015, caused the road surface to collapse near a disused underground tunnel, creating a large crater just meters from residential buildings. A few days before the collapse, residents had already reported signs of instability, including the collapse of a truck in the same area, an event that should have been a clear alarm bell. The collapse prompted the evacuation of approximately 400 people, leaving the neighborhood without water, gas, and electricity, and interrupting the Cumana railway line between Pianura and Soccavo. In addition to the material damage, the event highlights widespread structural issues related to the deterioration of underground water networks, the lack of timely maintenance, and the presence of unrecorded cavities beneath the urban fabric. The story highlights how the problem goes far beyond the single sinkhole, reflecting a systemic vulnerability that involves the management of underground infrastructure, public safety and urban resilience. [22]

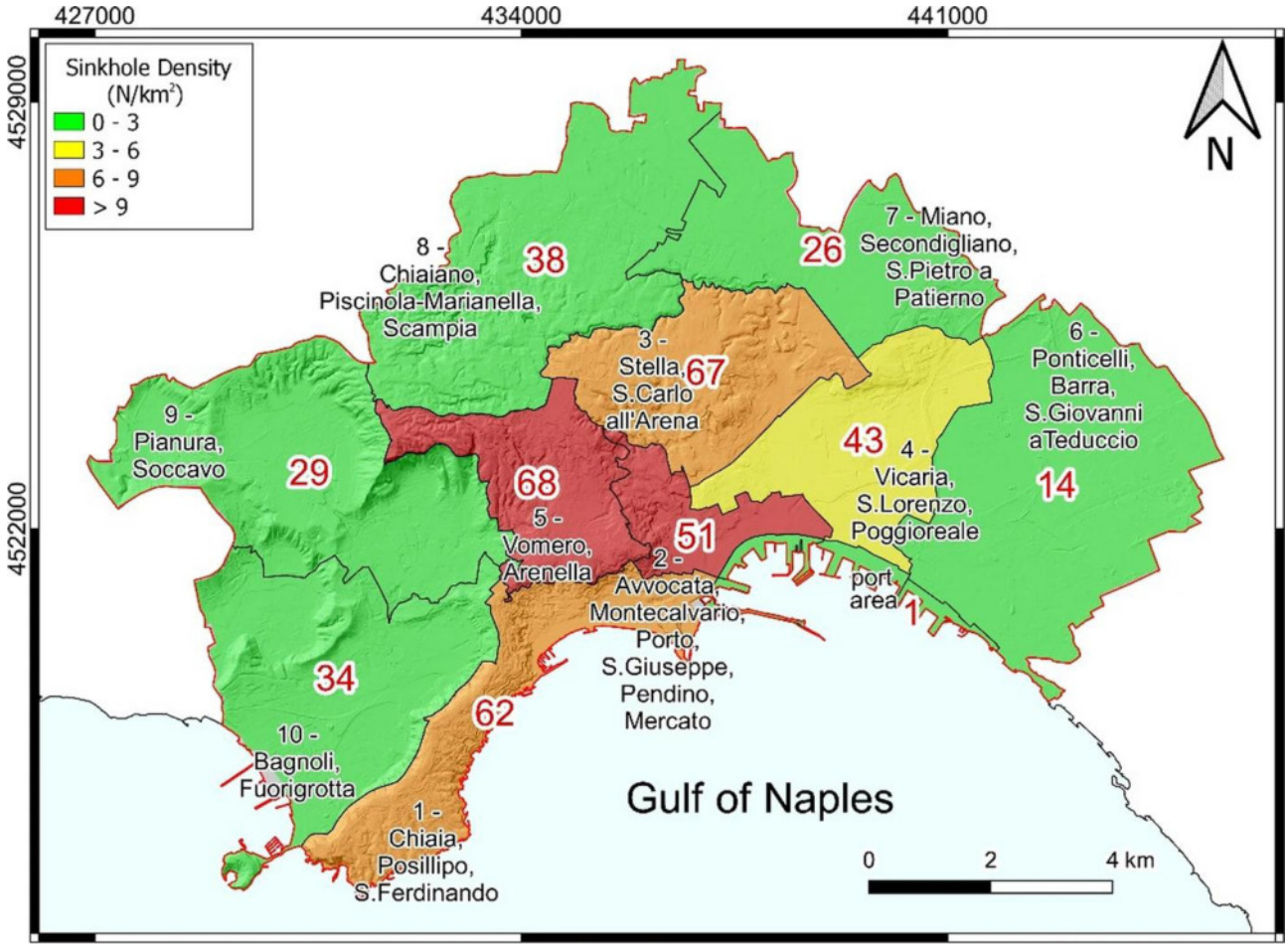

Figure 9. Reproduced from Tufano et al. (2022) [21]. Average distribution of sinkhole density in the various municipalities of Naples. The areas with the highest density of events are concentrated in the central sector of the city, particularly in the Vomero, Arenella, Avvocata, Montecalvario, Porto, S. Giuseppe, Pendino, and Mercato districts, where the presence of cavities and underground tunnels favors the formation of collapses. Peripheral areas, such as Ponticelli, Barra, Pianura, Bagnoli, etc., show a lower density.

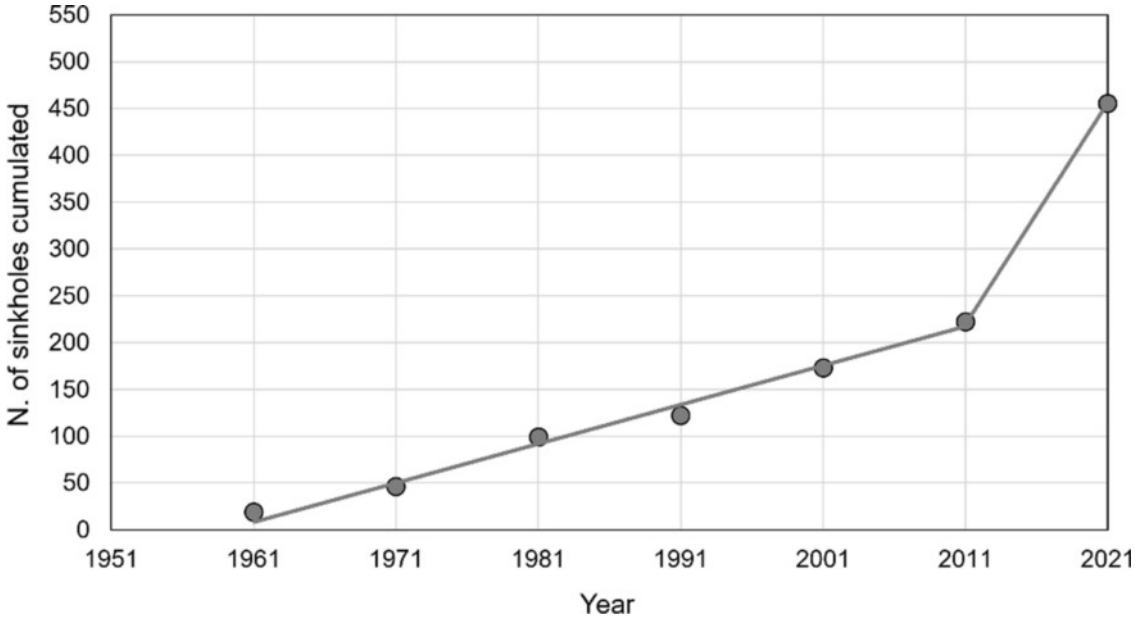

Figure 10. Reproduced from Tufano et al. (2022) [21]. Cumulative trend of the number of anthropogenic sinkholes recorded in the city of Naples between 1961 and 2021. The graph shows a progressive and constant growth over time, with a marked increase in recent years, which reflects both an actual increase in events and an improvement in detection and reporting activities.

Figure 11. The sinkhole on Via Campanile, in the Pianura neighborhood of Naples. Image source: ANSA (2015), "Voragine a Napoli, causa rottura condotta". [22]

In light of the studies analyzed, hydrogeological instability phenomena caused by water leaks in underground pipes are becoming increasingly significant in urban contexts, where high infrastructure density and the age of water networks amplify the negative effects of infiltration and land subsidence. In recent years, several cities have recorded an increasing number of land subsidences and collapses attributable to leaks in underground utilities; however, only a fraction of these events have been the subject of in-depth scientific analysis, while many remain documented only through journalistic surveys or local reports.

Experimental and numerical research shows how localized leaks can trigger processes of internal erosion, cavity formation, and, in the most severe cases, surface collapse, with serious implications for public safety, traffic flow, and infrastructure stability.

Specifically, literature highlights various triggers for leak-driven hydrogeological instability:

- the position, size and geometry of the defect (upper, lateral or lower part of the pipeline), which influence the infiltration path and the direction of the erosive cavity;
- the leak rate and flow velocity, which control the transition from simple infiltration to accelerated erosion or fluidization;
- the internal pressure of the pipelines and the variations in the hydraulic load, which determine the erosive force of the flow;
- the presence of external loads, such as traffic or vibrations, which amplify the deformations of the ground;
- the grain size, the content of fine particles and the stratigraphy of the soils that affect the resistance to erosion and the speed of collapse; In particular, profiles consisting of sandy clay, limestone and bedrock (SC–LS–BR) as well as those soils with a high amount of poorly cohesive fine particles were found to be the most vulnerable to the development of cavities;
- the depth of the pipeline and therefore the thickness of the covering layer, which influences the propagation time of the cavity towards the surface;
- the presence of anthropogenic cavities or disused tunnels, which create preferential paths for erosion;
- fluctuations in the water table and variations in pore pressure induced by leaks;
- pipes that have non-exceptional mechanical resistance as well as their poor maintenance;
- the absence of preventive monitoring systems;
- cyclic mechanical or thermal stresses, which contribute to the progressive deformation of joints and the formation of micro-cracks;
- the aging of the materials and the corrosion of the pipes, which increase the probability of localized failures;
- the interaction between networks of underground services (water, sewage, rainwater), which can amplify or accelerate collapse phenomena.
- intense and concentrated rainfall over time, which accelerates saturation and reduces the stability of surface soils;

Overall, these elements define a multidisciplinary cognitive framework useful for the construction of predictive models and mitigation strategies. [8]

However, as will be highlighted in the following paragraph, mitigation strategies for these phenomena do not yet have a truly integrated planning approach between hydraulic, geotechnical and urban planning components and, consequently, are not yet concretely considered and applied in urban land management policies.

1.2. Gaps in the literature

The previous section presented a review of the current state of knowledge and practice regarding the problem of hydrogeological instability caused by water leaks from underground pipelines. This section, however, aims to briefly analyze the current state of mitigation strategies for this risk, examining the main scientific and technical contributions. This will help us understand how network and land managers currently address the problem from a perspective of preventing and reducing the impacts of such leaks.

To outline the knowledge and operational gaps, it is necessary to identify the main disciplinary areas in which mitigation action can be implemented. These can be grouped into two main domains:

- the hydraulic area, which concerns the management, control and monitoring of water distribution networks;
- the geotechnical-geological field, relating to the behavior of soils, their stability and the monitoring of subsoil conditions.

The hydraulic sector is the most easily monitored and can be addressed promptly, as water leaks are the primary trigger for internal erosion and collapse. Early detection and localization of leaks can therefore be an effective mitigation tool, reducing response times and limiting the progression of damage.

In contrast, the geotechnical field presents greater monitoring challenges, as ground instability tends to manifest itself in the absence of surface evidence until the most advanced stages of the subsidence process. In this context, mitigation actions are most effective during the design and construction phases of the works, for example through proper subsoil modeling, the choice of backfill materials, and a trench configuration that reduces the vulnerability of the surrounding soils. During the operational phase, however, geotechnical interventions are more limited, making the contribution of hydraulic monitoring and maintenance practices crucial.

For the above reasons, the following bibliographic review will focus primarily on the hydraulic sector. Given the advanced age of many networks and the high level of leaks still present, it is now more urgent to develop tools and strategies capable of predicting and mitigating future hydrogeological instability caused by water losses. From this perspective, it is essential to promote an integrated and predictive approach, aimed at more carefully identifying the potentially most critical leaks and taking preventive action in urban areas characterized by a higher exposure value.

In this context, leak location plays a crucial role in mitigating instability and, even more so, maximizing the accuracy of leak location in areas at greatest risk of instability. Achieving this objective, however, requires zoning the risk of hydrogeological instability associated with leaks, thus directing monitoring activities toward the most vulnerable sectors of the network.

The required approach is therefore highly multidisciplinary, requiring collaboration between hydraulic, geotechnical, and land-use planning expertise. This complexity, combined with the limited understanding of the phenomenon and the fragmented nature of available experience, means that the scientific literature still lacks truly integrated methodologies capable of organically linking optimal leak location with landslide risk zoning.

In this direction, this work proposes the optimization of the positioning of pressure sensors as a connecting element, intended as a strategy to maximize the ability to identify and locate leaks based on coefficients representative of the territorial risk.

Therefore, a review of current leak detection and localization techniques, as well as sensor optimization methods developed in the literature to increase the efficiency and coverage of hydraulic monitoring, will be presented below. This analysis will allow to assess the degree of maturity of current knowledge and to highlight the distance that still separates research from a truly integrated approach for the mitigation of the risk of hydrogeological instability due to leaks in urban water networks.

The growing need for efficient and sustainable management of water resources requires network managers to adopt integrated strategies that include infrastructure redevelopment interventions, scheduled maintenance and pressure control [23], accompanied by advanced monitoring and diagnostic systems for early detection of leaks.

In this context, optimal sensor location plays a crucial role: improving the ability to detect and locate leaks not only reduces water waste and operating costs, but also, as seen, helps prevent ground instability caused by infiltration, particularly in urban areas with high exposure value.

Therefore, Optimal Sensor Placement (OSP) represents one of the main engineering challenges for the efficient management of water networks. It is no coincidence that many works exist regarding the optimization of sensor positioning, but in the following only a few are reported in an evolutionary and representative key, in order to understand the state of the art in view of the next steps of the thesis.

Casillas et al. (2013) introduced one of the first systematic approaches for optimizing the placement of sensors for leak detection in water distribution networks. The problem was formulated as a nonlinear integer optimization, with the aim of minimizing the number of non-isolatable leaks based on isolability criteria derived from sensitivity analysis. To address the computational complexity of the problem, the authors employed a genetic algorithm (GA), capable of identifying near-optimal solutions with a reduced computational load compared to exhaustive methods. The model, validated on two test networks, demonstrated robustness and flexibility, also thanks to the inclusion of distance, time horizon, and leak size criteria. [24]

The accuracy of leak location in a water network strongly depends on both the sensor placement and the uncertainties associated with measurements and water demands. In this perspective, Steffelbauer and Fuchs-Hanusch (2016) developed a methodology that integrates different sources of uncertainty within the optimal sensor placement (OSP) problem for model-based leak location. The method, tested on different network configurations and for various numbers of sensors, allows to evaluate how the presence of uncertainties modifies the quality of the location. The authors also derive a cost-benefit function that describes the relationship between the number of sensors and detection accuracy, observing that it follows a power law. The results highlight that including uncertainties leads to different sensor configurations compared to ideal scenarios and requires a greater number of devices to maintain the same localization performance. [25]

To improve the reliability of leak location in complex water networks, Cugueró-Escofet et al. (2017) developed a pressure sensor placement optimization methodology based on a relaxed isolation index. The approach considers realistic factors such as acceptable location distance and similarity of leak behavior between neighboring nodes, thus reducing identification errors and diagnostic ambiguities. Applied to two District Metered Areas of the Barcelona network, the method showed a significant improvement in leak isolation capacity. [26]

To address the lack of complete leak information, Li et al. (2019) developed a semi-supervised sensor placement optimization strategy. The method combines a feature selection algorithm, Semi-

supervised Joint Mutual Information (semi-JMI), with fuzzy C-Means clustering, allowing for partial leak localization and reducing monitoring blind zones. Applied to two reference networks, the approach showed significant improvement in terms of localization accuracy and stability. [27]

To increase the robustness of monitoring systems, Hu et al. (2021) proposed a hierarchical optimization method for sensor placement for leak detection based on an improved algorithm compared to traditional joint entropy-based models. The approach progressively selects sensors by minimizing the information loss in the event of a single sensor failure, thus ensuring greater system reliability. Applied to the EPANET Net3 test network, the method showed a significant reduction in entropy loss and greater adaptability compared to conventional algorithms, improving the leak detection and identification capability even in malfunction scenarios. [28]

To improve the efficiency of real-time monitoring systems, Cheng and Li (2023) proposed a sensor placement optimization method that combines feature selection and graph signal processing, avoiding the dependence on calibrated hydraulic models. The method analyzes the importance and redundancy of nodes, reconstructing unmeasured pressures and improving localization accuracy and speed compared to traditional methods. [29]

To address the challenges related to the scale and uncertainties of cyclic water networks, van Gemert et al. (2025) developed a graph analysis-based sensor placement algorithm capable of ensuring structural observability even in the presence of uncertain parameters. Tests on EPANET networks, including L-Town, showed computation times of less than 0.1 s, highlighting high efficiency and scalability of the method. [30]

Recent significant experiences have also been recorded in the European and Italian context. In the context of continuous monitoring of water networks, Batzella, Ferrarese, and Malavasi (2024) analyzed an active method for detecting and localizing breakages based on sensitivity matrices and correlation analysis. The study focuses on the influence of sensor sensitivity on the effectiveness of localization and proposes a strategy to identify the most effective positions of a predefined number of sensors, optimizing the ability to detect and isolate breakage events. [31]

In recent years, research has shown a growing interest in multi-objective and risk-based methods, in which the positioning of sensors is not only driven by hydraulic efficiency but also by criteria of social vulnerability and resilience.

The optimization of sensor placement for leak or break detection was addressed by Forconi et al. (2017) through a risk-based approach, which simultaneously considers the probability of non-detection, the potential impact, and the level of exposure associated with each leak event. Impact is related to the volume of water required, while exposure reflects the social, economic, or safety importance of water connections. Applied to a District Metered Area of the Harrogate network, the methodology showed how the inclusion of exposure can modify the ranking of optimal sensor locations, thus providing an installation criterion more consistent with the hydraulic, social, and economic needs of the system. [32]

To overcome the often unrealistic assumption that all nodes in a water network have the same relevance, Hu et al. (2022) proposed a multi-objective, risk-based approach for sensor placement optimization. The method introduces risk-based loss functions, which quantify the impact of a leak based on pressure, flow rate, and demand variations at network nodes. Through a multi-objective analysis, Pareto solutions are generated that balance detection effectiveness and risk reduction, which are subsequently evaluated through multi-criteria decision analysis. Applied to the C-Town network

model, the approach has been shown to provide more consistent sensor configurations with the detection priority of high-impact leaks, enabling more rational management in the presence of limited resources. [33]

With a view to increasing the resilience of urban water networks, Parajuli et al. (2025) developed an integrated framework for sensor placement and classification of different types of faults. The approach combines recursive feature selection via Random Forest (RF–RFE) with an Autoencoder–Random Forest (AE–RF) model capable of distinguishing leaks, physical attacks, and cyber attacks. Applied to the C-Town network, simulated with EPANET-CPA and WNTR, the method achieved accuracies up to 0.99, demonstrating its effectiveness in detecting and classifying faults in complex scenarios. [34]

These contributions, although oriented towards hydraulic efficiency, lay the foundations for an evolution towards integration with the assessment of the risk induced by the malfunctioning of the networks with respect to their impact on the external environment.

Overall, the literature on OSP shows steady progress towards smarter, more adaptive, resilient and less socially impactful methodologies but a significant gap remains: there is still a lack of an integrated framework that relates the optimization of pressure sensors with the zoning of the risk of hydrogeological instability from leaks.

At the moment, only the contribution of Medio et al. (2024) [35], whose work is part of this thesis, begins to connect the positioning of sensors with indicators of the risk of hydrogeological instability from leaks, paving the way for a multidisciplinary vision that combines hydraulic, geotechnical and environmental aspects.

This methodological gap represents an important research opportunity, especially in light of the current needs for urban resilience and risk management in complex infrastructure contexts.

1.3. Objective and structure of the thesis

The main objective of this research is to develop an innovative framework for mitigating the risk of hydrogeological disruption caused by leaks (HDL) in urban water distribution networks.

The study aims to design an optimization methodology for the positioning of pressure sensors, capable of maximizing the accuracy in the location of leaks and, at the same time, giving priority to urban areas characterized by a higher exposure value and, in a broader sense, risk from HDL.

The proposed framework integrates several technical components into a coherent multidisciplinary system:

- a hydraulic modeling that allows to simulate the hydraulic behavior of the network in the presence of different leak scenarios and to generate the data necessary for the construction of leak location algorithms;
- a risk zoning phase, developed through the combination of hazard, vulnerability and exposure factors;
- the definition of an objective function that integrates the accuracy of localization of leak events with the spatial distribution of hydrogeological risk;
- the use of evolutionary optimization algorithms to guide the process of positioning sensors according to risk-related priority criteria.

The approach is designed as an operational and decision support tool for local administrations and water network operators, combining the efficiency of hydraulic monitoring with the prevention of soil instability.

By identifying the most critical areas of the network, the framework enables targeted and rational monitoring, laying the foundation for more effective prevention precisely where the costs and social and infrastructural complications would be greatest in the event of hydrogeological instability induced by leaks.

The thesis is structured as follows:

- Chapter 2 is dedicated to materials and methods and illustrates the general structure of the proposed methodology, describing its main technical components. It first discusses the software used, such as EPANET for hydraulic modeling and the Python libraries for numerical implementation, followed by leak modeling, which is useful for characterizing the different simulated failure scenarios. Next, the risk zoning criteria and the objective function adopted to integrate accuracy and risk are illustrated. The chapter also includes an overview of the main leak location algorithms, both machine learning-based (e.g., Decision Tree) and deterministic (e.g., Cosine Similarity). The final section delves into evolutionary optimization algorithms, focusing on the custom Genetic Algorithm and the Pymoo framework, used to find optimal sensor placement.
- Chapter 3 is dedicated to case studies and consists of applying the proposed methodology to the analyzed networks. The characteristics of the *Real Network 1* and *L-Town* networks are presented, followed by a description of the simulation phases, parameter selection, testing procedures, and, finally, the results obtained. The *Real Network 1* constitutes the first testbed for validating the framework's fundamental concepts, while the *L-Town* network allows for verifying the model's methodological maturity and scalability on a benchmark network widely validated in the literature.
- Chapter 4 presents the conclusions and a critical reflection on the results. The main evidences that emerged, the application potential of the framework and future developments are discussed.

2. Materials and Methods

This chapter describes the methodology developed to identify the optimal set of pressure sensors capable of minimizing the risk of hydrogeological disruption induced by leaks (HDL) in water distribution networks (WDN) in the most critical areas.

To demonstrate the framework's effectiveness and modularity, the methodology is applied to two distinct case studies: *Real Network 1* first, and the *L-Town* network thereafter. While the overarching methodological structure remains consistent, different techniques are implemented across the two case studies. This approach serves to highlight the framework's adaptability to different network scales, urban contexts, modeling techniques, and levels of complexity.

2.1. Proposed Methodology and Workflow

As illustrated in the summary flowchart (Figure 12), the process is structured in four main integrated phases, moving from territorial risk analysis to the final optimization of the monitoring network.

Phase 1: Risk Assessment and Zoning

The methodological process begins with the Zoning of the network as a function of HDL risk. Rather than adopting a purely hydraulic approach, this framework establishes risk assessment as the fundamental prerequisite for monitoring. As detailed in Section 2.2, risk is quantified through the classic paradigm: $R = H \cdot V \cdot E$ (Hazard, Vulnerability, and Exposure). This phase allows the optimization to be "risk-oriented," prioritizing areas where a leak could cause the most severe structural and socio-economic damage. The zoning is achieved by integrating topological network data with land-use information and a detailed analysis of the Exposure (E) component, which accounts for the qualitative value of the involved areas, population density, and the strategic importance of the road network. In the second case study, the analysis is further expanded by developing a more in-depth assessment of the Hazard (H), correlating intrinsic factors (such as pipe diameters) with operational factors (such as operating pressures). Regarding the Vulnerability (V) component, in both case studies it is considered spatially constant due to the high uncertainty and lack of high-resolution stratigraphic data. By linking these elements, each node and pipe is assigned a specific risk level (Figure 12, Step 1).

Phase 2: Hydraulic Data Generation

Once the risk assessment has been developed, or in parallel with it, it is necessary to generate a robust dataset of pressure variations corresponding to different leak scenarios. Due to the scarcity of real-world leak data, this phase (described in Section 2.3) leverages a multi-software integration to create a synthetic benchmark. The core hydraulic engine of EPANET 2.2 is interfaced with Python-based toolkits such as EPyT and WNTR. The simulation methodology evolved from the application of the emitter coefficient via EPyT in the first case study, to the implementation of specific leak functions in WNTR in the second case study. These instruments allow for a physically-based modeling of leaks along the network, providing the necessary data for training localization models and evaluating sensor performance in a physically-consistent, pressure-dependent demand (PDD) environment (Figure 12, Step 2).

Phase 3: Leak Localization Modeling

The third phase involves the development, training, and testing of leak localization models designed to pinpoint a leak's location from sensor data. As explored in Section 2.4, the methodology implements two distinct computational strategies to demonstrate the framework's flexibility:

- A supervised data-driven approach using Decision Tree (DT) classifiers. This model is trained to learn the relationship between pressure patterns and leak locations, enabling localization without requiring an explicit physical model during the operational phase.
- A model-based approach based on the Sensitivity Matrix and Cosine Similarity, which identifies leaks by comparing observed pressure residuals with theoretical signatures derived from the hydraulic model.

The localization performance, expressed in terms of accuracy or hydraulic distance, serves as the primary metric to evaluate the effectiveness of different sensor configurations in the subsequent optimization stage (Figure 12, Step 3).

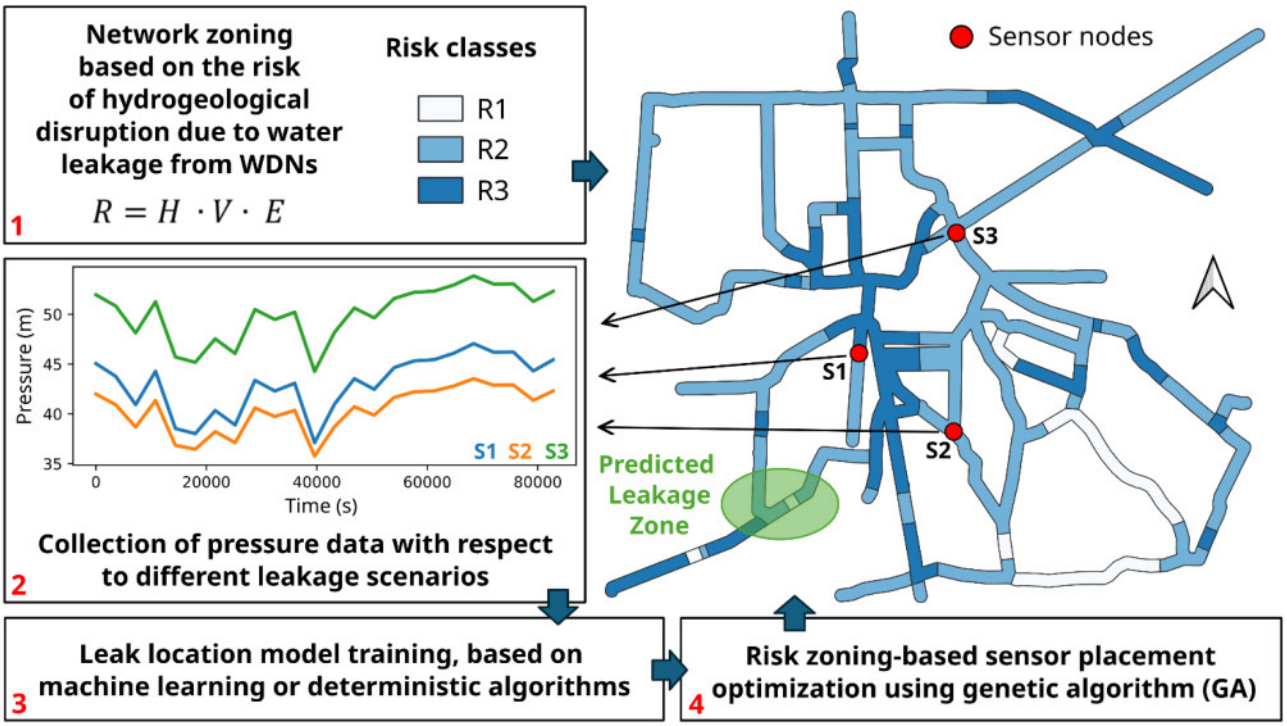

Figure 12. Summary of the proposed methodology. The process integrates four main phases: risk zoning (R) for hydrogeological disruption due to leaks (H = hazard, V = vulnerability, E = exposure), generation of hydraulic data, training of the leak localization model using machine learning or deterministic algorithms, and optimization of sensor positioning. Risk classes range from the lowest level (R1) to the highest (R3). Adapted from Medio et al. (2024) [35].

Phase 4: Risk-Oriented Sensor Optimization

The final stage (detailed in Section 2.5) synthesizes all previous elements into an optimization framework based on Genetic Algorithms (GA). The GA explores the vast space of possible sensor configurations to identify the set that maximizes monitoring effectiveness. The innovation of this approach lies in the "risk-oriented" fitness function: by using the risk weights (W_z) defined in Phase 1, the algorithm prioritizes configurations that ensure the highest localization accuracy specifically in high-risk zones, while maintaining an acceptable level of global accuracy across the entire network. Two different optimization implementations are presented: a customized GA focused on weighted accuracy (A_w) and an approach based on the Pymoo framework focused on minimizing the weighted

minimum hydraulic distance (D_w), demonstrating the consistency and replicability of the entire operational flow. (Figure 12, Step 4).

2.2. Risk Evaluation of Hydrogeological Disruption Due to Water Leaks

As discussed in the introductory chapter, water losses in urban environments can generate ground instability phenomena with different degrees of severity.

The hazard of the event is mainly determined by the hydraulic and structural characteristics of the pipeline, such as the operating pressure, the flow rate, the diameter, the material and the state of degradation, which influence both the probability of rupture and the amount of water potentially dispersed in the ground.

The severity of the damage that can result is instead conditioned by the local geotechnical conditions and the strategic or infrastructural value of the area involved, elements that contribute to defining the vulnerability and exposure of the system respectively.

In this context, risk assessment is not an ancillary analysis but constitutes the methodological prerequisite for guiding the entire process of optimizing the positioning of the sensors.

Locating leaks with the utmost accuracy in areas with the highest vulnerability or exposure value makes it possible to mitigate the risk of hydrogeological disruption in a targeted manner, reducing the potential socio-economic impact associated with land subsidence or infrastructural damage.

The goal is not only to identify the weak points of the network, but to establish a priority of intervention based on the spatial distribution of risk, so as to make monitoring and maintenance more effective and sustainable.

In this perspective, the concept of risk is adopted here as a key component of the optimization framework. It therefore represents the combination of three main factors (hazard, vulnerability and exposure) [36].

This formulation allows us to describe in a concise but effective way the probability and severity of the potential damage resulting from a localized leak event.

The quantification of risk must therefore take into account the following three key factors [35]:

- *Hazard*, *H*: the likelihood of occurrence of the dangerous event, i.e., the hydrogeological disruption caused by a non-detected leak with a given magnitude at a given location;
- *Vulnerability*, *V*: the expected degree of damage due to the impact of the hazardous event (hydrogeological disruption due to leakage) on the system (soil, urban infrastructures and human elements);
- *Exposure*, *E*: the socio-economic importance of goods, structures, and infrastructures, as well as the presence of people in the at-risk area.

This formulation, although simple, encapsulates the cardinal principle of modern risk analysis: there is no risk without the concomitance of a potentially dangerous event, a system vulnerable to suffering damage and a value exposed to the consequences. All this can be expressed through the following product [36]:

$$R = H \cdot V \cdot E \quad (1)$$

This formulation makes it possible to combine the hydraulic component (linked to the probability of leak) with the geotechnical and territorial component (related to the potential effects on soil and infrastructure). The goal is not only to identify the areas most prone to network failures, but above all to recognize the areas where a possible leak would have the most serious consequences in terms of safety, economic damage and urban impact.

In the present study, the risk is graded on a scale of N_R levels, ranging from R_1 (minimal risk) to R_{NR} (maximum risk). Let $\Omega = \{1,2,\dots,N_N\}$ be the set of junction nodes (and, similarly, links) of the water distribution network (WDN), the function $R(i) = R_j$, with $i \in \Omega$ and $j \in \{1,2,\dots,N_R\}$, associates each node (or link) with the corresponding risk level of risk R_j.

Starting from the definition of Equation (1), the assessment of the risk associated with instability phenomena induced by water leaks must be based on an integrated analysis of the characteristics of the distribution network (WDN), the geotechnical behavior of the soil, the mechanisms of cavity formation and the presence of exposed elements.

In recent years, several studies have addressed the issue of sinkhole risk in urban environments, proposing different approaches for risk zoning mapping.

Among these, Bianchini et al. (2023) conducted research in the plain of Guidonia–Bagni di Tivoli (Rome), an area characterized by unstable travertine deposits and frequent sinkhole episodes. The authors developed a susceptibility and risk map by integrating geological, lithological and hydrogeological factors with InSAR satellite data, estimating susceptibility using a machine learning model based on the Maximum Entropy (MaxEnt) algorithm. The subsequent combination with the components of vulnerability and exposure has made it possible to generate an urban risk map, highlighting that the areas at highest risk (about 2% of the total area) coincide with the consolidated and infrastructural urban fabric, where 27% of buildings and 5% of the road network fall into the high or very high risk classes. This approach showed the effectiveness of predictive models in supporting spatial planning and geological risk management, although water leaks from underground pipelines were not explicitly considered as a triggering factor. [37]

A further methodological contribution is that of Zhang et al. (2024), who developed an innovative approach based on Grey System Theory for mapping urban sinkhole susceptibility and risk in the city of Shenzhen, China. The methodology integrated geological, meteorological and infrastructure data with subsidence information obtained from InSAR, using Grey Relation Analysis (GRA) to estimate the relative weight of each factor. The susceptibility map thus obtained was then validated and combined with information on the urban transport system to produce a specific risk map for road infrastructure, revealing critical areas where preventive and targeted monitoring measures can be taken. This integrated approach has demonstrated the ability to accurately identify urban areas at risk, even in contexts of high geotechnical and infrastructural complexity, but similarly to the case of Bianchini he does not consider leaks of water pipelines as one of the main factors on which to intervene to mitigate the spread of sinkholes. [38]

It is precisely in this context that the present work is inserted, which aims to fill this gap, introducing for the first time a zoning of the risk from water leaks aimed at mitigating urban instability.

The developed methodology integrates the concepts of hazard (H), vulnerability (V) and exposure (E) with hydraulic modeling and leak localization, allowing to identify the areas where the probability of causing damage is greater as a function of the behavior of the network and the elements exposed to disruption.

This zoning is also directly linked to the optimization phase of the positioning of the pressure sensors, so as to maximize the ability to locate leaks precisely in the areas with the highest hydrogeological risk.

After defining the general HDL risk report, it is useful to examine in more detail the characteristics of the vulnerability, hazard and exposure components in the context of water leaks in underground pipelines and how the sinkhole literature has already treated them.

- *Vulnerability* is often complex and subject to uncertainty in its assessment process, since for a given adverse event, it depends substantially on the resistance of the different elements exposed.
  In the case at hand, the vulnerability assessment requires detailed information on the geological, hydrogeological and in particular geotechnical characteristics of the areas exposed to the leak, since factors such as soil cohesion, permeability, grain size, presence of artificial fills or pre-existing cavities strongly influence the response of the soil to the flow of pressurized water. However, this level of detail can only be obtained through accurate identification and stratigraphic characterization of the soils. This type of survey, as a rule, concerns limited areas and cannot be considered representative of the vulnerability of the entire urban water network.
  Furthermore, as already highlighted by Medio et al. [35], the technical and stratigraphic information of the subsoil provided by the various bodies, while constituting a useful reference base, does not have sufficient resolution and continuity for detailed zoning, making it difficult to directly transpose this information into vulnerability indicators.
  Recent studies confirm the difficulty in quantifying vulnerability on an urban scale. For example, Bianchini et al. (2023) [37] have shown that, even in areas subject to natural sinkholes or induced by human activities, the identification of the most vulnerable areas requires the integration of multiple factors, including lithology, travertine thickness, groundwater depth and land use.
  As an example, Intrieri et al. (2023) [39], in the study on the municipality of Camaiore, considered vulnerability as a function of structural strength and speed of collapse, attributing a value of 1 to all structures potentially subject to large sudden sinkholes, as such events would be sufficient to destroy them or make them unusable. This approach highlights how vulnerability can also be modeled through indirect parameters, linked to the scale and dynamics of the phenomenon.
  In the absence of precise data, it is therefore instinctive to adopt a uniform vulnerability value for the area under consideration. This simplification pushes us to focus the analysis on the other components of risk, i.e. hazard and exposure, as they are normally accompanied by a greater wealth of information on which to make more detailed assessments.
- *Hazard,* in the context of hydrogeological instability induced by water leaks, is a function of the structural, operational and environmental conditions that influence its onset. It is therefore strongly linked to the probability of network failure and constitutes the dynamic component of the risk, as it describes the expected frequency and intensity of the phenomenon.
  Several studies have shown that hazard is influenced by multiple concomitant factors: pipe material, age and state of degradation, pressure and hydraulic fluctuations, characteristics of the surrounding soil (humidity, pH, chemical aggressiveness), as well as climatic and operating conditions.
  Among the most recent contributions, Barton et al. (2019) [40] have highlighted how predictive modeling of ruptures must consider not only infrastructural data but also

environmental and pedological data, since the interaction between pipe and soil (e.g. changes in moisture and soil shrinkage) can amplify mechanical stresses on pipelines, increasing the probability of rupture (Figures 13, 14, and 15) . The authors also emphasize the importance of correlating statistical models with the operational experience of field technicians, as errors in data collection or fault classification can introduce significant biases into analyses.

An in-depth study on the causes of corrosion, the main factor in the degradation of metal pipes, has been proposed by Farh et al. (2023) [41]. In this study, the authors classify the factors that accelerate the corrosive process into three main categories (Figure 16): environmental factors (temperature, soil type, etc.), pipe-related factors (age, material, etc.), and operational factors (pressure, quality, etc.). The analysis by Fuzzy Analytical Hierarchy Process showed that operational factors have the greatest weight in the probability of failure, followed by environmental ones, confirming the importance of operating conditions in the assessment of hazard.

Similarly, Muddassir et al. (2024) [42] developed a probabilistic approach to identify the most significant factors in pipeline failure, applying a combination of Bayes theorem and likelihood feature selection. The results show that materials such as galvanized iron and polyethylene are more vulnerable.

A further contribution comes from Philip and Aljassmi (2020) [43], who in a review on the deterioration of water networks highlight the importance of accurate predictive models for the estimation of the residual useful life of pipes. The authors highlight how the scarcity of reliable data and the difficulty in calibrating degradation models still represent one of the main barriers to predicting the real risk of failure.

Finally, recent studies such as Sinaei et al. (2025) [44] and Wéber et al. (2020, 2022) [45, 46] have adopted integrated approaches, combining probability of failure (POF) models with social, economic and environmental indicators to quantify the consequence of failure (COF). In particular, Sinaei et al. apply a Weibull model to estimate the hazard as a function of the age of the pipeline, while Wéber proposes an approach based on the theory of complex networks, in which the probability of failure of the segments is combined with the network topology and with the distribution of consumption to assess the overall vulnerability of the system.

Overall, it emerges that hazard cannot be considered a static parameter, but the result of multi-factorial interactions between environmental conditions, construction characteristics and hydraulic behavior.

- *Exposure*, for the type of risk under consideration, generally represents the most immediately quantifiable component and describes the socioeconomic and infrastructural value of assets and areas potentially affected by instability phenomena induced by water leaks. It measures the extent of potential damage not in terms of probability of occurrence (such as hazard) nor the physical vulnerability of materials but rather based on the presence and value of the exposed assets, be they homes, infrastructure, economic activities, or the resident population.

  In the case of urban water networks, exposure depends on the topological configuration of the network and therefore on the density of vulnerable elements in the areas adjacent to the pipelines, such as buildings, roads, transport lines or strategic infrastructures (hospitals, schools, energy networks). Knowledge of the topology of the network and the analysis of the economic and social value of the assets present in its surroundings are generally sufficient to estimate this component [35].

  A solid methodological example is offered by Bianchini et al. [37], which construct exposure and vulnerability maps by integrating: land registers and geo-topographic databases

(buildings, infrastructures, services); real estate values OMI (Osservatorio Mercato Immobiliare – Real Estate Market Observatory) for the monetization of the exposure of residential buildings (€/m²) and other land use classes; the CORINE classification to associate exposure values to the different categories of land use. The result is a map that allows the potential damage in areas adjacent to the network to be coherently weighed, distinguishing, for example, residential areas (higher OMI values) from agricultural or semi-natural areas (much lower values).

Therefore, the estimation of exposure is often related to land use and settlement density, so it follows that the use of appropriate datasets and censuses allows to assign to each node or section of pipeline a weight proportional to the type of area crossed (residential, industrial, agricultural, infrastructural, etc.).

In light of the above, it should be noted that the exposure component tends to exhibit strong spatial discontinuities: densely built-up areas or those with a high concentration of critical infrastructure generate risk peaks even in the presence of relatively low levels of risk, and vice versa. This implies that, in monitoring strategies, priorities should not be defined solely based on network conditions, but also on the potential socioeconomic impact of a potential leak.

Figure 13. Reproduced from Barton et al. (2019) [40]. Main categories of factors influencing the failure of water pipelines. The diagram distinguishes between intrinsic factors of the pipeline (e.g. age, diameter, joints, coatings, manufacturing defects), environmental factors (such as corrosion, hydrogeological conditions, soil type and climate) and operational factors related to network management (pressure, thermal variations, maintenance operations). The interaction between these elements determines the probability and the failure mode of the pipeline.

In the present study, as well as in the literature [39], in the absence of detailed and homogeneous information on the spatial distribution of some risk component, in order to arrive at a general assessment, a numerical simplification is adopted, which may consist, for example, in considering one or more uniform components over the entire analysis territory. This hypothesis allows us to express, for example, the risk as:

- $R = c_1 \cdot E$
  where $c_1 = H \cdot V$ represents a spatially uniform constant. In practical terms, it assumes:
  $H = const.$ and $V = const.$

or:

- $R = c_2 \cdot H \cdot E$
  with $V = const.$

This assumption, although simplifying, allows us to focus attention on the spatial distribution of the exposure (and possibly of the hazard) and its relative weight in the optimal location of the sensors, thus laying the foundations for a risk zoning aimed at the most urbanistically sensitive areas.

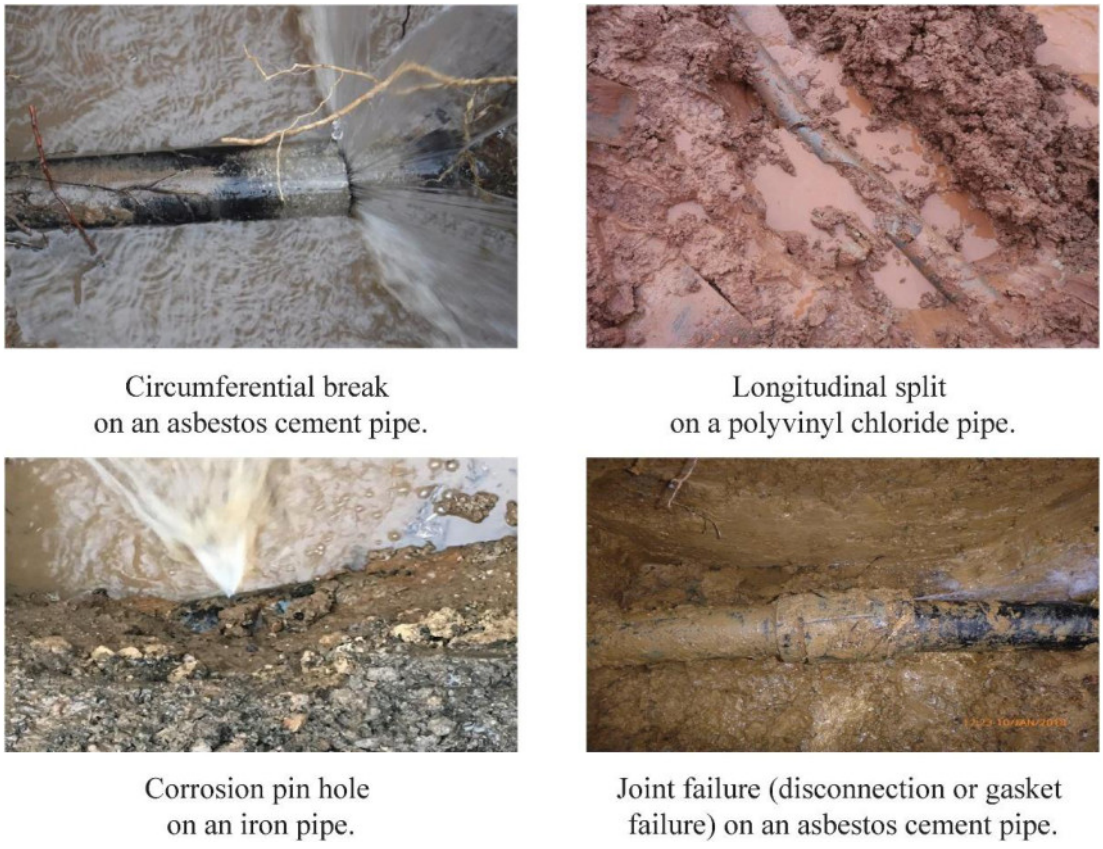

Figure 14. Reproduced from Barton et al. (2019) [40] and adapted in turn by Anglian Water (2018) [47]. Examples of different types of failures observed in water pipes.

Figure 15. Reproduced from Barton et al. (2019) [40]. Examples of corrosion and chemical degradation observed in cast iron pipes. The image shows two typical deterioration phenomena: graphitization (left), in which the metal matrix is progressively replaced by residual graphite, and pitting corrosion (right), characterized by the formation of localized cavities that weaken the wall of the pipe to the point of perforation.

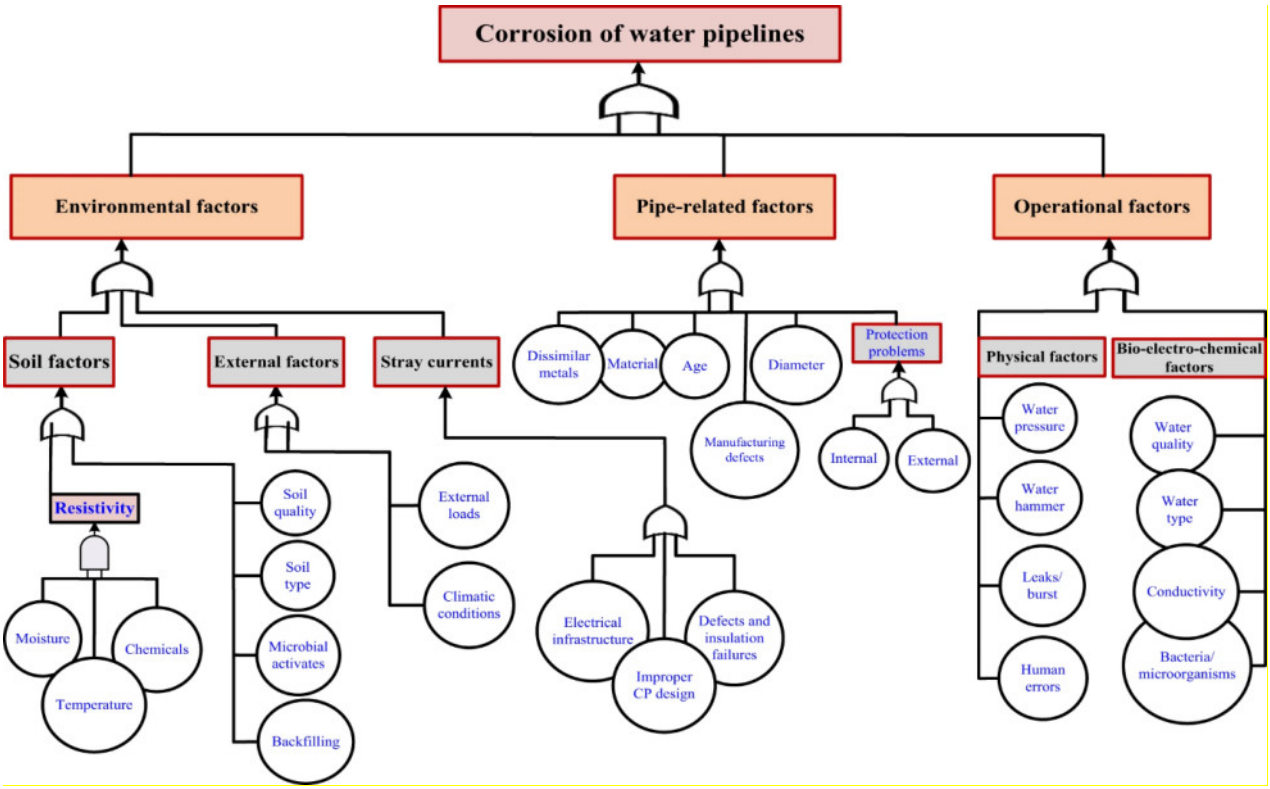

Figure 16. Reproduced from Farh et al. (2023) [41]. An overall summary of the main causes of corrosion in water pipes, divided into environmental, intrinsic pipe and operational factors. The classification, proposed by Farh et al. (2023), highlights how corrosion derives from the interaction between hydro-geochemical conditions of the soil, construction and material properties of the pipelines, and how the network operates. This conceptual approach is in line with the one presented by Barton et al. (2019), which adopts a similar subdivision to describe the mechanisms of degradation of water infrastructure.

2.3. Hydraulic data generation (or collection) for different leak scenarios

This section describes the software and libraries used for the generation of leak scenarios and for the consequent collection of hydraulic data used in the analysis carried out.

In addition, some theoretical and operational concepts are introduced that are fundamental for the correct understanding of the research work.

Since the work includes two case studies, it should be noted that in the first one only the EPyT library was used, while in the second, in order to refine the modeling of the loss phenomenon, thanks to a greater knowledge and awareness of the available tools, the EPyT and WNTR libraries were used together.

2.3.1. EPANET

This study uses synthetic hydraulic data generated with the open-source software EPANET 2.2 [48], developed by the U.S. Environmental Protection Agency (EPA).

EPANET is one of the most widely used reference tools worldwide for the simulation of the hydraulic and qualitative behavior of pressurized water distribution networks.

The software makes it possible to analyze the variation over time of quantities such as the flow rate in the pipelines, the pressure in the nodes, the level in the tanks and the concentration of any dissolved substances in the water.

Since its release in the 1990s, EPANET has established itself as a reference platform for the research sector and for the operational management of water networks, thanks to its reliability, modular structure and completely open-source nature, which facilitates its integration with other analysis and simulation tools.

Version 2.2, adopted in this work, introduces important improvements compared to the previous ones: greater numerical stability, more efficient error handling and compatibility with Unicode file formats, which expands its use in international research contexts.

The EPANET calculation model is based on the simultaneous solution of the continuity and energy equations for each node and pipeline of the grid.

The hydraulic behavior of a pressurized water distribution system can be represented through a set of nonlinear equations that guarantee the conservation of mass and energy along the network.

The continuity equation, written for each node i expresses the balance of incoming and outgoing flows, taking into account demand and any local inflows or disbursements:

$$\sum_{j \in \text{pipes connected to the node } i} Q_j = D_i \tag{2}$$

where Q_j represents the flow rate in the pipeline j and D_i the demand assigned to the node i.

This formulation ensures that, for each node, the algebraic sum of the flows entering and leaving is zero, in compliance with the principle of conservation of mass.

For each pipe that connects the nodes i and k, the energy equation describes the total load difference between the two extremes, including frictional and localized losses:

$$H_i - H_k = h_f(Q) + h_m \tag{3}$$

where H_i and H_k are the piezometric heights at the ends of the pipe, $h_f(Q)$ represents the pressure drop distributed along the pipeline and h_m the localized loss associated with bends, valves, or fittings.

The distributed pressure drops can be calculated according to different empirical formulations. In the default model, EPANET uses the Hazen–Williams equation. Alternatively, the user can also select the Darcy–Weisbach or Chezy–Manning formulations (see Table 1).

Table 1. Pressure drop formulas for full complement pipelines (for headloss in feet and flow rate in CFS), reproduced from the EPANET 2.2 User Manual, U.S. EPA, 2020 [48].

Formula	Resistance Coefficient (A)	Flow Exponent (B)
Hazen-Williams	$4.727C^{-1.852}d^{-4.871}L$	1.852
Darcy-Weisbach	$0.0252\,f(\epsilon, d, q)d^{-5}L$	2
Chezy-Manning	$4.66\,n^2\,d^{-5.33}\,L$	2

with:

- C = Hazen-Williams roughness coefficient;
- ϵ = Darcy-Weisbach roughness coefficient (ft);
- f = friction factor (dependent on ϵ, d, and q);
- n = Manning roughness coefficient;
- d = pipe diameter (ft);
- L = pipe length (ft);
- q = flow rate (cfs).

In addition, EPANET 2.2 allowed for the implementation of a Pressure Dependent Demand (PDD) model in the present study. To describe this pressure dependency, the software employs the power law expression developed by Wagner et al. (1988) [49], which calculates the actual demand d_j delivered at a node based on the available pressure head P_j (defined as the hydraulic head H_j minus the node's elevation Z_j):

$$d_j = \begin{cases} D_j & if\ P_j \geq P_{ser} \\ D_j \cdot \left(\dfrac{P_j - P_{min}}{P_{ser} - P_{min}}\right)^{1/e} & if\ P_{min} < P_j < P_{ser} \\ 0 & if\ P_j \leq P_{min} \end{cases} \tag{4}$$

where D_j is the full normal demand at node j when the pressure P_j reaches or surpasses the service limit P_{ser}, while P_{min} denotes the threshold below which the demand is zero. Furthermore, the exponent $1/e$ characterizes the pressure-dependency of the flow and is typically assumed to be 0.5 to simulate the hydraulic behavior of discharge through an orifice.

The overall system of nonlinear equations is solved through the so-called Gradient Algorithm, a modified version of the Newton–Raphson method, developed by Todini and Pilati (1988) [50] and subsequently integrated into the EPANET code.

This algorithm allows to separate the continuity equations (associated with the nodes) from the energy equations (associated with the pipes), reducing the size of the system to be solved and improving numerical stability even for large networks.

Thanks to this formulation, EPANET guarantees rapid convergence and remarkable computational efficiency, while maintaining high accuracy in the estimation of hydraulic loads and pressures.

In addition to the strictly hydraulic aspects, the software makes it possible to represent networks with complex topologies, to manage time-varying demands through daily or weekly patterns and to simulate water quality processes, such as residual chlorine concentration.

The openness of the source code has also fostered the emergence of numerous programming libraries, such as EPyT and WNTR, which further extend its capabilities and allow hydraulic simulations to be integrated with statistical analysis and machine learning tools, particularly in the Python language.

2.3.2. EPyT - EPANET Python Toolkit

Among the most recent and advanced libraries for interfacing with EPANET, an important role is played by the EPANET-Python Toolkit (EPyT), developed at the KIOS Research and Innovation Center of Excellence of the University of Cyprus [51].

EPyT provides a comprehensive, object-oriented Python interface that integrates directly with EPANET's hydraulic and water quality calculation engine, allowing you to run simulations, modify network model parameters, and manage results in a simple and automated way.

The toolkit was created with the aim of simplifying and standardizing the use of the EPANET library within the research community on "Smart Water Networks", reducing the level of computer skills required to interact with the native API.

As reported by the authors, the toolkit was designed to address four main needs of Smart Water Networks (SWN) research:

- Provide a standardized framework that allows researchers to develop, share and replicate scientific methodologies in a coherent and interoperable way;
- Reduce the time and complexity required to establish the connection with the EPANET libraries, making it easier to run custom simulations;
- To provide code templates to extend the functionality of EPANET and promote the dissemination of good practices in open science and reproducible research;
- Ensure a uniform data structure between Python and MATLAB, facilitating the translation of code developed in academia (often in MATLAB) to industrial or Python-based application contexts.

Unlike simpler wrappers or generic links to EPANET libraries, EPyT provides direct access to all of the more than 500 functions and parameters of the original engine, while maintaining the same nomenclature and data structure as the EPANET-MATLAB Toolkit (EMT), from which it inherits the design philosophy.

This bidirectional compatibility allows you to easily integrate methodologies developed in MATLAB into Python environments, facilitating code reuse and collaboration between research groups.

From an operational point of view, EPyT allows you to import hydraulic models in *.inp* format, perform hydraulic and water quality simulations, modify network parameters (such as diameters, roughness, demand patterns or pump curves) in real time and automatically save the results in analyzable format.

The main object of the toolkit, called epanet, contains the topological information of the network and the calculation functions.

For example, to load and simulate the L-TOWN network, all you need to do is:

```
from epyt import epanet
G = epanet('L-TOWN.inp') # Load the L-TOWN network model
H = G.getComputedHydraulicTimeSeries() # Run hydraulic simulation
Q = G.getComputedQualityTimeSeries() # Run water quality simulation
```

Once the G-object is created, the user can visualize the network, modify the properties of nodes and pipes, run multiple simulations, and graph the temporal evolution of pressures or flows.

EPyT includes built-in functions for the automatic visualization and management of results, as well as the ability to export datasets directly in a format compatible with NumPy, Pandas or Matplotlib, making it a particularly versatile tool for statistical analysis, machine learning models and optimization.

In the present study, in particular in the first case, EPyT has been used to generate a synthetic dataset of pressures containing two leakage scenarios.

This approach made it possible to ensure uniformity of simulation conditions, reproducibility of results and scalability towards more complex networks, in line with the objectives of research on smart water systems.

2.3.3. WNTR (Water Network Tool for Resilience)

The Water Network Tool for Resilience (WNTR) is an open-source Python package, developed by the U.S. Environmental Protection Agency (EPA) in collaboration with Sandia National Laboratories, with the aim of simulating and analyzing the resilience of water distribution networks. [Klise et al. (2017) [52]; Klise et al. (2018) [53]; Klise et al. (2020) [54]]

The software is fully compatible with EPANET and allows you to represent in detail the hydraulic behavior of the network in the presence of critical events such as leaks, power failures, contamination, fires or pipe breaks.

Compared to EPANET, WNTR significantly extends simulation capabilities, introducing tools to assess resilience, model complex failure scenarios, and test mitigation and recovery strategies.

The Python interface and integration with scientific libraries such as NumPy, SciPy, Pandas and Matplotlib allow you to analyze the structure of the network, manage large volumes of hydraulic data and visualize the results through high-resolution graphs and animations.

Leak modeling is one of the most innovative aspects of WNTR, as it allows the phenomenon to be represented in a physically coherent way, overcoming the limitations of the simplified emitter-based model of EPANET.

In WNTR, leaks are simulated through a dedicated model implemented in the WNTRSimulator, which calculates the leakage flow rate as a direct function of the pressure at the node or at the point of pipe failure.

In particular, the loss model adopted in WNTR is based on the formulation proposed by Crowl and Louvar (2001) [55], representing a rewriting and characterization of the well-known Bernoulli/Torricelli relationship, in which the loss rate Q_L (m³/s) is expressed as:

$$Q_L = C_d \, A \sqrt{2gh} \tag{5}$$

where:

- C_d is the exhaust coefficient (dimensionless), typically equal to 0.75 in turbulent conditions;
- A is the equivalent area of the hole (m²);
- g is the acceleration of gravity (m/s²);
- h is the manometric pressure expressed as piezometric height (m).

The model also includes a loss exponent of 0.5, which realistically represents the behavior of large losses on metal pipes, a value consistent with what has been observed in the experimental literature. [56]

Unlike EPANET, where the emitter coefficient must be empirically calibrated, in WNTR the physical parameters of the leak, such as hole diameter, equivalent area or local pressure, are explicitly defined, allowing for more direct and transparent modeling of the phenomenon.

Leaks can be assigned to:

- nodes or tanks, representing a localized emission;
- pipelines, by inserting an intermediate node generated with the split_pipe() function.

This way it is possible to locate the leak in a specific position in the pipeline and control its activation over time.

The add_leak() method allows you to define the area of the hole, the start and end time of the event (in seconds) and more. A typical example of an implementation is as follows:

```
wn = wntr.morph.split_pipe(wn, '123', '123_B', '123_leak_node')
leak_node = wn.get_node('123_leak_node')
leak_node.add_leak(wn, area=0.05, start_time=2*3600, end_time=12*3600)
```

This function automatically adds a time control to the simulation, activating and deactivating the loss over the specified period.

From a physical point of view, WNTR considers the leak as an additional pressure-dependent demand: in low pressure conditions the leakage flow rate is reduced, making the model consistent with the Pressure Dependent Demand (PDD) [49] approach adopted in the simulator.

In this way, it is possible to represent dynamic and realistic scenarios, which include not only the leak, but also any pressure changes, repairs or subsequent restarts of the simulation (pause and restart).

With this formulation, WNTR enables more accurate analysis of the hydraulic effects of leaks and generates consistent and reproducible simulated datasets that are useful for sensitivity analysis, automatic leak location, and network resilience assessment applications.

2.4. Leak localization models: model-based, data-driven, and hybrid approaches

The localization of leaks in water distribution networks is one of the most complex and crucial activities in the management of aqueduct systems. The main difficulty derives from the intrinsic uncertainties of the hydraulic model (related to the variability of demand, the roughness of the pipelines, the network schematization, and others) and the limited number of sensors available in real-world contexts. Under such conditions, it is reasonable to consider a leak 'localized' when it is possible to narrow down the set of candidate pipes (or nodes, for more simplified localization problems) to a sufficiently restricted subset, within which the source of the failure is highly probable to be found. The exact location can then be determined through field surveys, conducted with acoustic, geophysical, thermographic, or other specialized instruments.

As highlighted by El-Zahab and Zayed (2019) [57], research in the field of leak detection has undergone a significant evolution, with the transition from purely experimental methodologies to intelligent systems of continuous monitoring. The authors propose a three-step classification, Identify, Localize, Pinpoint (ILP), which describes the logical process of leak detection:

- Identification, aimed at determining the presence of an anomaly and distinguishing it from other causes (network manoeuvres, demand variations or instrumental disturbances);
- Location, aimed at delimiting the sector or group of pipelines in which the leak occurred;
- Precise identification, which allows you to define the exact location of the fault for subsequent inspection or repair.

In this context, two main classes of detection systems are distinguished [57]:

- static systems, based on fixed sensors that monitor pressures or acoustic signals continuously and transmit data to a control center. They allow early identification, but can be subject to false alarms due to hydraulic noise;
- dynamic systems, which make use of mobile instruments, such as geophones, penetration radar (GPR), infrared thermal cameras or robotic devices (smart-balls), for targeted and high-precision inspections.

Static systems are more effective in the Identify and Localize phases, while dynamic systems find their maximum application in the Pinpoint phase, i.e. in the exact determination of the leak. [57]

A significant contribution to the systematization of leak detection and localization methodologies is provided by Romero-Ben et al. (2023) [58], who propose a structured classification based on the degree of dependence of the method on the hydraulic model of the network (Figure 17).

According to the authors, localization techniques can be divided into three broad categories: model-based, mixed model-based/data-driven, and data-driven, each of which has specific advantages and limitations.

- *Model-based methods*: these are based on the use of a detailed hydraulic model, which can also be implemented in simulation software (e.g. EPANET), considered as a high-fidelity representation of the hydraulic behavior of the network.
  The localization of the leak is carried out by comparing the real data (pressures or flow rates) with those simulated by the model, analyzing the systematic deviations.
  These approaches are very accurate in the presence of well-calibrated models but suffer from the difficulty of defining and updating hydraulic parameters, as well as sensitivity to measurement errors.

In the literature, this category includes methods based on inverse transients, sensitivity analysis, Bayesian inference, and fuzzy logic, including contributions by Pudar & Liggett (1992) [59], Casillas et al. (2014) [60], Pérez et al (2014) [61], Zhang & Wang (2011) [62], Sanz et al. (2012) [63], and Bicik et al. (2011) [64].

- *Mixed model-based/data-driven methods* represent an evolution of traditional model-based approaches and were created with the aim of mitigating the difficulties related to the direct use of the hydraulic model in the leak location process.
  These difficulties include the complexity in the selection and calibration of mathematical models, the diversity and structural complexity of water networks, and the presence of modeling errors due to uncertainties about nodal demands or instrumental noise. [58]
  To overcome these limitations, the hydraulic model is mainly used in offline phases, for example for the generation of synthetic datasets, which are then used for training data-driven models such as machine learning. The operational localization phase, on the other hand, is generally entrusted to data-driven algorithms, such as machine learning, capable of analyzing the pressure/flow time series acquired by the network sensors in real time.
  This creates a synergy between the physical robustness of the hydraulic model and the adaptive flexibility of data-driven methods, improving the system's ability to operate in variable conditions and in the presence of noise.
  Within this category, the literature proposes several strategies for integrating modeling and machine learning, including artificial neural networks, Support Vector Machines (SVMs), neuro-fuzzy systems, and deep learning networks, each of which represents a different way of combining physical knowledge and artificial intelligence.
  A representative example is provided by Sun et al. (2019) [65], who propose a leak localization method. The model involves two phases: in the first, two machine learning classifiers, Linear Discriminant Analysis (LDA) and Neural Network (NNET), are used to estimate the probability that each node in the network is the site of a leak; in the second, a Bayesian temporal reasoning analysis dynamically updates these probabilities over time, improving the accuracy of the localization.
  Tested on the Hanoi DMA reference network, the method achieved accuracies of more than 80% in the best cases, demonstrating how the integration of hydraulic modeling and probabilistic learning allows effective localization even with a limited number of sensors and a reduced computational load.
  A further hybrid methodology is that of Ares-Milián et al. (2021) [66], which combine a multiclass SVM classifier with an inverse problem solved using a variant of the differential evolution algorithm. In a first step, the SVM is trained on the simulated data from the hydraulic model to reduce the search space, identifying the most likely area where the leak is located. In the second step, point localization is formulated as an inverse problem and solved by the Topological Differential Evolution (TDE) algorithm. The results, obtained considering demand and noise uncertainty scenarios (5–15% of nominal demand), show greater accuracy and robustness than purely data-driven or purely deterministic methods, highlighting the advantage of the sequential combination of learning and optimization.
  Wachla et al. (2015) [67] propose a hybrid approach based on adaptive neuro-fuzzy systems, applied to the localization of leaks in networks divided into predefined areas. Each area is associated with a neuro-fuzzy classifier that receives the flow or pressure residues as input, i.e. the deviations between the values predicted by the predictive models and those actually measured. The outputs of the classifiers indicate the probability of leakage in the corresponding area, allowing the area affected by the fault to be identified.

Instead, Li et al. (2022) [68] propose a methodology based on deep neural networks of the ResNet (Residual Network) type for the accurate localization of leaks in water distribution networks. The model integrates a classification process with a regression process to estimate the exact locations of leaks. In addition, thanks to a multi-supervision mechanism introduced in the regression process, the convergence of the model is accelerated, making it more performing. The results confirm the high reliability of the model even in complex networks. These approaches have proven to be particularly effective in combining the physical interpretability of the model with the adaptive flexibility of machine learning algorithms.

- *Data-driven methods.* As pointed out by Romero-Ben et al. (2023) [58], data-driven methods represent the most recent and dynamic category among strategies for identifying and locating leaks in water networks. Unlike model-based and mixed approaches, these methods rely solely on data measured by devices installed in the network, typically pressure, flow or level sensors, without requiring a hydraulic model of the system. In other words, the analysis and decision-making process is entirely based on extracting knowledge from the observed data, without the need for information from simulations or calibrated models, which are not always available or updated from water utilities.

  These approaches reduce reliance on hydraulic models, improving the adaptability of monitoring systems and their applicability in real-world operational contexts.

  In particular, the literature distinguishes different families of data-driven methodologies such as: approaches based on statistical analysis, machine learning techniques, topological interpolation and analysis methods and advanced probabilistic approaches.

  A first group of data-driven methods is based on the statistical analysis of pressure and flow data to identify anomalous variations indicative of leaks. A significant example is provided by Romano et al. (2017) [69], who develop a leak location system called Leakage Location System (LLS) based on Statistical Process Control (SPC). The method analyzes nighttime data collected every one minute by pressure loggers installed in water districts (DMAs), identifying statistically significant deviations from the normal behavior of the network. The system has been tested in over 130 events showing a remarkable ability to reduce the research area and operating costs, as well as operational and environmental benefits related to the reduction of lost water volumes.

  Within the same category are also approaches that exploit machine learning or data mining techniques to model the behavior of the network based on empirical data.

  Laucelli et al. (2016) [70] propose a model based on the Evolutionary Polynomial Regression (EPR) paradigm, which allows to reproduce and predict the hydraulic behavior of a network using real-time data from low-cost sensors.

  Applied to a real network of a DMA, the method demonstrated high accuracy in the reconstruction of hydraulic behavior and in the detection of flow anomalies, suggesting the possibility of integrating these models into early warning systems. This approach highlights the potential of evolutionary and regressive techniques for predictive analysis of hydraulic data, without the need for physical models.

  A significant evolution of data-driven methods is represented by models that integrate measured data and topological information from the network, to improve the spatial localization capability of the leak.

  Alves et al. (2021) [71] propose a robust approach based on pressure and flow measurements combined with system topology, which uses a reduced-order model to estimate "no-leak" pressures at monitored nodes.

The pressure residues, calculated as the difference between the estimated values and those actually measured, are then correlated with the topological incidence of each node through a leak probability index.
This index is based on an incidence factor that takes into account the most likely hydraulic path between the tanks, sensors and nodes potentially affected by leakage. The model proposed by Alves and colleagues stands out for its ability to integrate the network structure with real-time data, ensuring reliable results without the need for a complete hydraulic model.
Data-driven methods therefore also include probabilistic methods, which are based on forecasting water demand and analyzing deviations from expected values.
Hutton and Kapelan (2015) [72] present a Bayesian Demand Forecasting approach, which probabilistically predicts future demand under normal conditions and assesses the probability that an observed measurement will be anomalous.
Leaks or breakages are then identified as events with low probability compared to the expected behavior.
Applied to a UK water network, the method showed good performance in detecting in real time leaks of more than 5% of the average daily flow during the night, demonstrating the versatility of the Bayesian paradigm for detecting anomalies in demand time series.
Taken together, data-driven methods are characterized by the ability to analyze large amounts of operational data and recognize anomalous patterns without the need for an explicit description of the system.
Their effectiveness depends to a large extent on the quality and frequency of the data available, as well as the presence of sensors distributed in strategic locations in the network.
Thanks to advances in data analysis techniques and sensors, these approaches now represent one of the most promising directions for automatic leak detection, particularly in real networks where hydraulic models are incomplete or poorly updated.

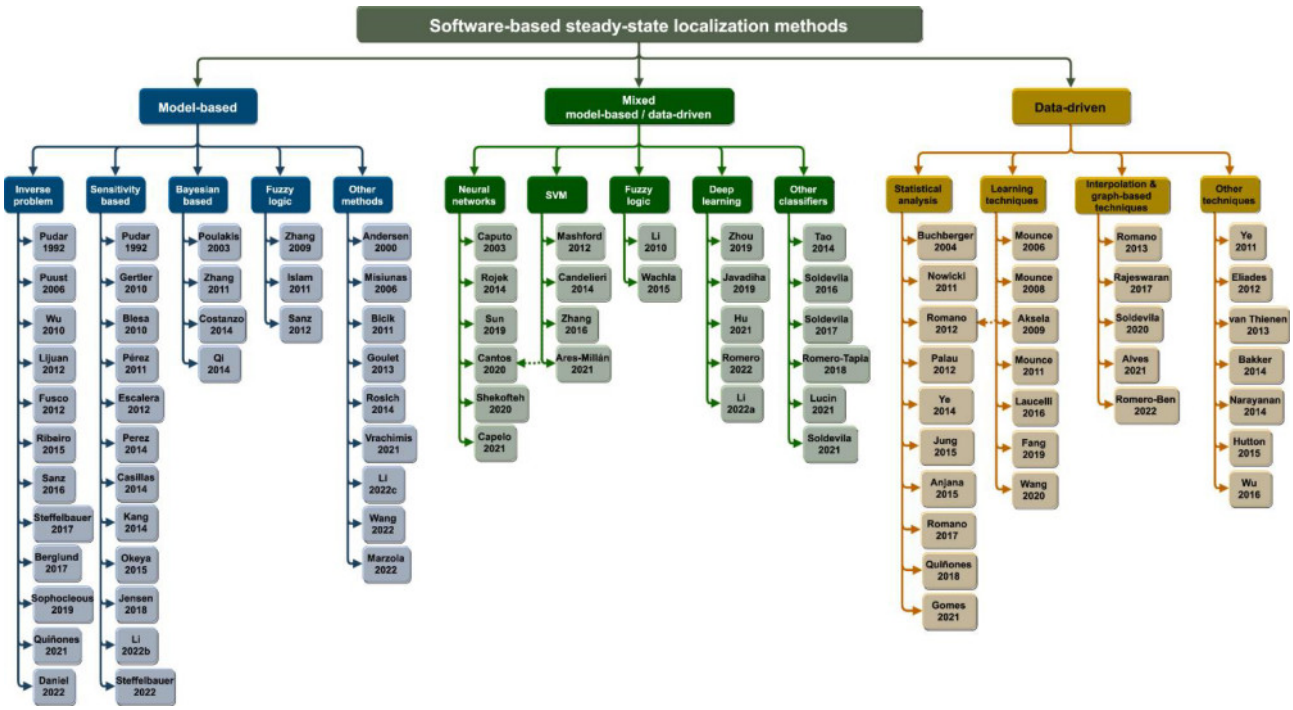

Figure 17. Reproduced from Romero-Ben et al. (2023) [58]. Hierarchical classification of software-based methods of leak localization, divided into model-based, hybrid (model-based/data-driven) and data-driven approaches.

In the present work, the leak location phase plays a functional and demonstrative role within a broader process, aimed at mitigating the hydrogeological risk induced by leaks in urban water networks.

The main objective of the research is in fact to build an integrated methodological structure, capable of linking the different phases (from the generation and/or analysis of leak scenarios to the localization and subsequent optimization of the positioning of the sensors, as well as to the mitigation of the risk, which in turn is based on its own zoning process) in a coherent framework, modular and replicable.

Although in the present work the different steps are demonstrated through simulated data, the entire process is designed to be extensible to real data or acquired in real time, without changing its logic or operational sequence.

In this perspective, the localization methods implemented do not represent cutting-edge solutions in terms of algorithmic accuracy or complexity but have been chosen for their simplicity and conceptual transparency, so as not to weigh down the overall narrative of the proposed method.

Since the process is conceived in a modular form, the replacement of these algorithms with more sophisticated methodologies is not only possible, but desirable, especially in applications on real or real-time pressure datasets, where the complexity of hydraulic behavior and measurement uncertainties require more advanced analysis tools.

The goal is therefore to illustrate the consistency and replicability of the entire operational flow, rather than maximizing the accuracy of the single localization phase.

It should also be noted that in this study the leak detection phase is not treated, as it is assumed that the system comes into operation after the identification of an anomaly or leak already detected through one of the numerous leak detection methods widely documented in the literature (based, for example, on acoustic signals, transient analysis, mass balances or statistical methodologies).

The focus is therefore placed exclusively on the localization phase, which allows the most probable leak zone within the network to be estimated starting from the observed pressure variations.

For reasons of consistency and controllability of the results, as well as for the difficulties encountered in obtaining real data, all the data used in this experimental phase were generated through hydraulic simulations in EPANET, considering different leak configurations.

These data have a demonstration purpose, aimed at testing the operation of the integrated process, and therefore should not be interpreted in absolute terms of accuracy or generalizability of the results, since they depend on the hypotheses and simulation conditions adopted.

For the same reasons, no competitive comparisons (battle-style) were made between different localization methods, nor was a performance classification of the results obtained conducted.

The aim of this research is not to propose a new localization technique capable of surpassing others in the literature, but to integrate existing methodologies within a modular and coherent process, aimed at showing how leak localization can be functionally connected to the optimization phase of sensor positioning and risk zoning.

In the context of this research, two distinct approaches have been adopted for leak location, representative of the main methodological categories present in the literature:

- a supervised data-driven approach, based on a Decision Tree classifier, which learns from the pressure patterns associated with the leak scenarios generated with EPANET;
- a model-based approach, based on the sensitivity matrix and on the comparison by cosine similarity between the theoretical leak signatures and the observed pressure variations, also generated with EPANET.

Despite their simplicity, these methods allow to effectively represent the two main families of techniques for leak localization which are therefore included in the broader process of integration between hydraulic analysis, HDL risk analysis and optimization of sensor positioning, which constitutes the real innovative focus of this work, which will be shown in more exhaustive detail below.

2.4.1. Supervised data-driven approach: Decision Tree classifier

In recent decades, machine learning (ML) has taken a central role in scientific research and technological innovation, driven by the increasing availability of data and computing power, along with the evolution of new machine learning algorithms [73].

The idea of allowing a computer to "learn" abstract concepts from data, then applying them to situations not yet observed, is not recent: the first neural networks date back to the 50s, when they were introduced as mathematical models of the human brain [73]. Earlier approaches, such as Bayesian statistics and Markov chains, also reflected the insight that a system could adapt its predictions based on experience.

As Badillo et al. (2020) [73] point out, machine learning today sits between two "cultures" of statistical modeling, a conceptualization introduced by Breiman [74] in 2001:

- the first culture, model-based, provides for the explicit formulation of a mathematical model to describe the observed data, based on theoretical principles and interpretability of relationships;
- the second culture, algorithmic, aims to identify relationships and patterns in data without specifying a model a priori, delegating the discovery of underlying connections to the computer.

Machine learning is predominantly in the latter category: models are not imposed but learned from data through the optimization of objective functions. This makes them particularly powerful but, at the same time, less interpretable, the so-called black box problem. However, in recent years several interpretability techniques have been developed, which allow us to better understand the inner workings of complex models. [73]

As highlighted in a recent scholarly work, Spector (2024), the relationship between Artificial Intelligence (AI) and Data Science is characterized by a complex intersectional structure (Figure 18), with Machine Learning (ML) serving as the central methodological core for both fields. The distinct areas of AI encompass domains such as symbolic logic, knowledge representation, and robotic sensing, among others. Conversely, Data Science maintains a specific focus on statistics, operations research, data visualization, and so forth. [75]

While ML is a primary driver, a substantial overlap exists between AI and Data Science independent of it, dedicated to solving complex problems at scale. Within this shared space, the two disciplines

converge to address critical challenges in sectors such as privacy, security, governance, and healthcare, and more. [75]

This distinction, often overlooked in non-specialist literature, is important in this type of work.

Within machine learning, four main categories of learning are traditionally distinguished [76]:

- *Supervised learning*, in which the model is trained on known input-output pairs, to learn a predictive function;
- *Unsupervised learning*, aimed at identifying hidden structures or groupings in unlabeled data;
- *Semi-supervised learning*, which combines the two previous logics, partially exploiting labeled data;
- *Reinforcement learning*, where an agent learns to perform optimal actions by maximizing a reward function.

Sarker (2021) [76] highlights how these approaches constitute the core of Industry 4.0 techniques, finding application in areas ranging from cybersecurity to smart cities, from healthcare to precision agriculture. In all cases, the effectiveness of the model depends heavily on the nature of the data and the choice of the most appropriate learning method.

The author also emphasizes the importance of feature engineering and dimensional reduction (e.g. through PCA - Principal component analysis), crucial aspects for the construction of efficient data-driven systems. [76]

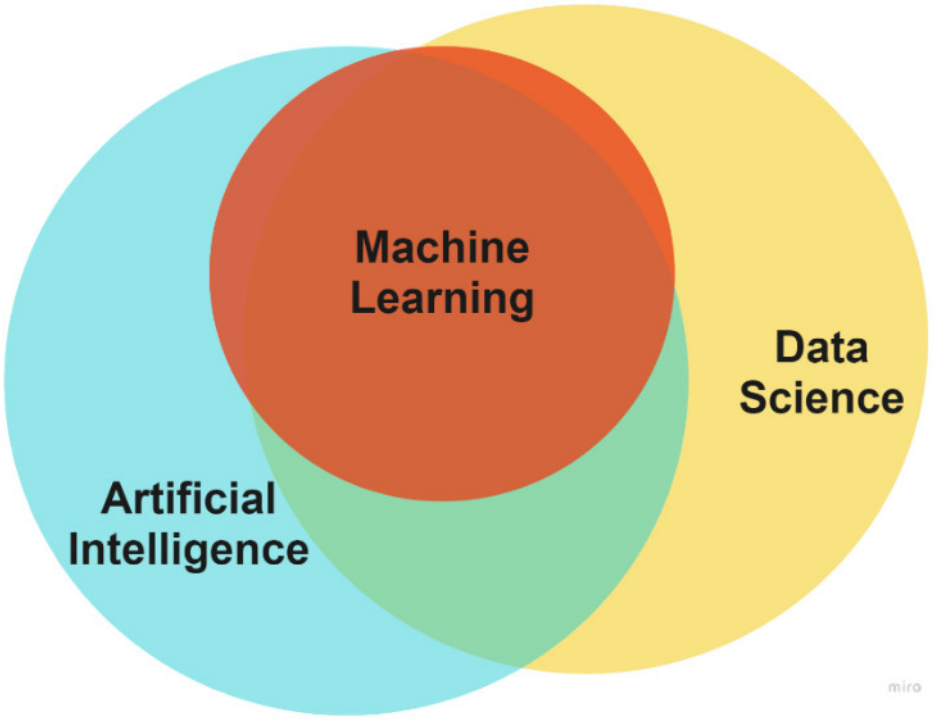

Figure 18. Reproduced from Spector. (2024). The conceptual intersection between Artificial Intelligence, Data Science, and Machine Learning. [75]

Overall, machine learning today represents a flexible and adaptive modeling paradigm, capable of learning autonomously from past experiences to make predictions or decisions on new data. Its use

spans a wide range of scientific and technological domains, offering a bridge between traditional statistical modeling and the latest data-driven AI approaches.

In the present work, the localization of leaks has been addressed, firstly, with a supervised data-driven approach based on Decision Tree (DT). The basic idea is simple and transparent: the model learns, starting from labeled examples, the relationship that links the pressure values recorded by the sensors (feature) to the position of the leak (label). In this case, the features are the pressures measured or simulated at the sensor nodes during each leak scenario, while the labels correspond to the nodes (or pipes) or clusters of nodes (or pipes) where the leak originates. Trained on these examples, the algorithm is able to associate new pressure observations with the most likely leak class.

While the decision tree is not the best in terms of accuracy or predictive capabilities, it was chosen for its conceptual simplicity, speed of training, and natural integration into a modular pipeline. These characteristics are perfectly consistent with the objectives of this work, whose aim is not to compare the performance between localization methods, but to demonstrate the validity of the entire methodological framework aimed at mitigating the risk from HDL.

The Decision Tree (DT) represents one of the most widespread methodologies of supervised machine learning. [Song & Lu. 2015 [77]; Blockeel et al. 2023 [78]]

This approach requires a training dataset containing both the features (input variables) and the corresponding labels (desired output). In the case in question, the labels identify the leak site node, while the associated features consist of the pressure values recorded by the sensor nodes in the presence of each leak.

The algorithm constructs a hierarchical tree model, in which the population of the data is recursively divided into more homogeneous subsets.

The process starts with a root node, which represents the entire dataset, and proceeds through a series of internal nodes and branches that reflect the different decision-making choices, until it reaches the leaves, which represent the final classes (i.e. localized leak scenarios).

As highlighted by Song & Lu (2015) [77], this structure allows to efficiently manage large and complex datasets without imposing parametric assumptions on the distribution of data, making the algorithm particularly suitable for nonlinear problems and problems with many related variables.

The model was trained using the *DecisionTreeClassifier* function of the scikit-learn library. [79]

In particular, the Gini criterion (criterion='gini') was used to evaluate the purity of the partitions: at each step, the algorithm selects the feature and threshold that maximize the reduction of impurity, obtaining the best separation between the classes.

The procedure continues until all leaves are pure, i.e. they contain only samples belonging to a single class, or when the minimum number of samples per leaf (min_samples_leaf = 2) is reached.

A more modern and important aspect should also be highlighted: their transparency compared to other machine learning algorithms, such as neural networks or deep learning models, which are considered "black boxes". Decision trees, in fact, are interpretable models (explainable models), i.e. they allow us to understand how and why the model has made a certain decision.

Once trained, the DT model is used to predict the location of the leak based on the pressure values contained in the test set.

For a given sensor set *P*, the localization accuracy $M_k(P)$ for a specific localization cluster Ω_k is defined as:

$$M_k(P) = N_{A_k}(P)/N_k \quad (6)$$

where:

- k is the index identifying the specific localization cluster being evaluated;
- Ω_k represents the k-th localization cluster, defined as a specific sub-area of the water distribution network (WDN) to which a leak is associated within the localization process, consisting of a set of neighboring nodes and/or pipes grouped based on their topological proximity and generally share common physical or hydraulic characteristics (e.g., elevation, pressure zones, districts, and so forth);
- $N_{A_k}(P)$ represents the number of correct predictions made by the sensor set *P* regarding the leaks originating from the nodes/pipes of the localization cluster Ω_k in the test dataset;
- N_k is the total number of leak scenarios generated by nodes/pipes in the same cluster Ω_k in the test dataset.

This measure quantifies the probability that the selected set of sensors *P* will correctly identify the leaks coming from the cluster Ω_k within the test dataset [35].

2.4.2. Model-based approach: Sensitivity matrix and cosine similarity comparison

Model-based methods are based on the construction of a mathematical or computational model that describes the behavior of a system and the relationships between its variables. Unlike data-driven approaches, which are limited to identifying empirical correlations, model-based approaches use structured assumptions to represent the mechanisms that generate data.

In model-based methods for leak location, a central role is played by the sensitivity matrix, a tool that quantifies the influence of each possible leak on the pressures detected in the network. This approach is based on the principle that, in a hydraulic system, the operating variables (such as pressure and flow rate) are mutually dependent, and a localized variation (e.g. a leak) is consistently reflected at multiple points in the network.

As reported by Hu et al. (2021) [80], the sensitivity matrix was introduced by Pudar and Liggett (1992) [81] and subsequently applied in a number of studies, including those by Pérez et al. (2009 [82], 2011 [83], 2014a [84], 2014b [85], 2016 [86]). It is built from a calibrated hydraulic model, generally developed in environments such as EPANET, and represents the pressure changes at the nodes in response to small changes in flow rate or demand associated with each element of the network. Each column in the matrix then describes the effect of a potential leak on all monitored nodes.

The analysis is based on the comparison between the pressures simulated by the model and those measured in the field: the difference, or residual pressure, indicates the deviation from the expected behavior. By comparing the residue vector with the columns of the sensitivity matrix, it is possible to identify the most probable location of the leak, as its "impact" on the system will be more similar to the theoretical signature calculated for that specific pipeline.

Despite the high theoretical effectiveness of this method, its accuracy strongly depends on the quality of the calibration and the management of uncertainties related to unknown questions, measurement noise and model approximations (Hu et al., 2021 [80]; Blesa & Pérez. 2018 [87]). For this reason, targeted improvements have been developed in the most recent studies [80].

In summary, the sensitivity matrix provides a quantitative representation of the link between leaks and pressure changes, constituting the starting point for localization methods based on vector similarity, such as those that employ the cosine similarity to compare simulated scenarios and real measurements.

From a mathematical point of view, we can represent the sensitivity matrix, *S*, as [81, 83]:

$$S = \begin{pmatrix} \frac{\partial p_1}{\partial f_1} & \cdots & \frac{\partial p_1}{\partial f_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial p_n}{\partial f_1} & \cdots & \frac{\partial p_n}{\partial f_n} \end{pmatrix} \tag{7}$$

where each element $\partial p_i/\partial f_j$ represents the change in pressure at the node due to a change in flow rate (or leak) associated with the element.ij.

The matrix thus constructed then describes the mutual dependence between the nodes of the network and the hydraulic variations induced by a local anomaly.

This formulation allows to identify the characteristic response of the system to an elementary perturbation, providing the basis for the generation of theoretical signatures useful for the localization of leaks. By comparing these signatures with the measured pressure residues, it is possible to estimate the pipeline most likely to leak.

From a methodological point of view, one of the most critical aspects in the matching processes between theoretical signatures and residues lies in the dependence on thresholds. Normally, in fact, it is necessary to establish a limit value that allows you to distinguish whether a pressure variation between model and measurement is to be considered significant. The choice of this threshold, however, is far from trivial: values that are too low generate numerous false positives, while values that are too high lead to a loss of sensitivity and therefore false negatives.

As described by Pérez et al. (2011) [83], the sensitivity-based analysis process involves a sequence of three distinct steps. First, as we have seen, we construct the sensitivity matrix S, which quantifies the change in pressure at each node of the network in response to a hypothetical localized leak on each pipeline. Since real water networks constitute multivariable, nonlinear and non-explicitly solvable systems, this matrix cannot be obtained analytically but is generated numerically by successive hydraulic simulations, introducing a unit leak in each node of the model.

Subsequently, the values of the matrix are normalized in order to make the responses of the different sensors comparable. In particular, each row of the matrix (corresponding to a sensor) is divided by its maximum value, thus obtaining the normalized sensitivity matrix, $\bar{S}$:

$$\bar{S} = \begin{pmatrix} \frac{s_{11}}{\sigma_1} & \cdots & \frac{s_{1n}}{\sigma_1} \\ \vdots & \ddots & \vdots \\ \frac{s_{n1}}{\sigma_n} & \cdots & \frac{s_{nn}}{\sigma_n} \end{pmatrix} \tag{8}$$

where all elements are between 0 and 1.

From this, the so-called leak signature matrix (or binarised sensitivity matrix) is obtained, which represents the significant relationships between leaks and pressure variations in a simplified form. Its construction requires the definition of a threshold: the elements of the normalized matrix above this value are assigned to 1 (relevant effect), while the lower ones are set to 0 (negligible effect).

The authors point out that the choice of this threshold is a crucial aspect, as it determines the final structure of the signature matrix. In particular, low thresholds produce an almost fully populated matrix, in which each node is affected by almost all leaks, while high thresholds make it almost diagonal, indicating that each sensor responds primarily to the nearest leak. This strong dependence on the threshold involves the need for a preventive optimization phase to identify the value that maximizes the localization capacity, while introducing greater computational complexity and sensitivity to operating conditions.

In the present study, a significant optimization phase is already envisaged, concerning the positioning of sensors in the network. Introducing additional optimization procedures, such as finding the optimal threshold for binarizing the sensitivity matrix, would therefore be inefficient. Furthermore, since the optimal configuration of the threshold can vary as the set of sensors considered varies, this would lead to a significant increase in computational complexity and overall calculation time.

To overcome this limitation, in this work we adopt a threshold-independent similarity measure, based on cosine similarity, which evaluates the directional coherence between the residual vector r and the sensitivity vectors s_N associated with the different loss scenarios.

As also reported in Bartkowska et al. (2024) [88], it is possible to write:

$$\cos(\theta) = \frac{r \cdot s_N}{|r|\,|s_N|} \tag{9}$$

where $|r|$ and $|s_N|$ represent the Euclidean norms of the respective carriers.

The value of $\cos(\theta)$, between -1 and 1, allows you to quantify the degree of directional alignment between the two signatures in the pressure space, regardless of the absolute intensity of the variations.

Therefore, the cosine similarity proves to be more compatible for the purposes of this work since it eliminates the need to calibrate empirical thresholds and maintains good performance even in the presence of noise and measurement uncertainties.

The adoption of this metric therefore offers a double advantage:

- on the one hand, it allows a direct and continuous comparison between theoretical signatures and residues without resorting to arbitrary thresholds;

- on the other hand, it significantly reduces the need for further optimization steps, which is particularly relevant in the context of this work, where the ultimate goal is to optimize the positioning of the sensors based on HDL risk zoning.

2.5. Risk-Oriented Optimization of Sensor Positioning Using Genetic Algorithms

The optimization process is a crucial step in the development of the framework that leads to leak location. In general terms, optimization can be defined as the systematic search for the best configuration of decision variables, such as to maximize or minimize an objective function, while respecting a set of physical or operational constraints.

In this research, the optimization problem has been formulated with the specific aim of identifying optimal sets of pressure sensors capable of maximizing the accuracy of leak localization under certain conditions, taking into account the hydraulic and topological characteristics of the network. In addition, a risk-oriented perspective has been adopted, which assigns a higher priority to the correct location of leaks that occur in areas characterized by a potentially higher impact, according to the risk zoning previously defined.

To solve the problem, different optimization strategies were tested, implemented and compared.

The first approach is based on an in-house developed genetic algorithm designed specifically for this application. This approach has allowed a high degree of control over evolutionary operators (selection, crossover, mutation, elitism) and the definition of specific constraints of the problem, such as the guarantee of uniqueness of sensors within each individual and the elimination of duplicate solutions between generations. This customised implementation proved to be particularly useful in the early stages of the research, as it made it possible to gain an in-depth understanding of the dynamics of the evolutionary process and to integrate specific assessment metrics, as well as objective functions that took into account the concept of risk.

The second approach, on the other hand, was based on the use of the Pymoo framework [89], a consolidated Python library for evolutionary and multi-objective optimization.

Among the various tools tested, Pymoo was the most suitable for the needs of this research. The reasons for this choice lie in its flexibility in defining constraints and objective functions, in the advanced management of evolutionary populations and in the possibility of customizing genetic operators in a similar way to what has been done in the code developed internally. In addition, the library's modular structure has simplified the integration of new valuation metrics, such as the risk-weighted average hydraulic minimum path.

Finally, both approaches, the one with the optimization algorithm developed in-house and the one based on Pymoo, will be presented in the rest of the work, as they constitute two fundamental stages of the research path developed during the PhD. In addition to representing two different levels of methodological complexity, they are applied to distinct case studies, allowing, from an overall evaluation perspective, to draw broader and more significant considerations on the effectiveness and generalizability of the proposed overall approach.

2.5.1. Characteristics of genetic algorithms

Genetic Algorithms (GAs) are research and optimization techniques inspired by the mechanisms of natural evolution and genetics. The idea stems from the pioneering studies of John H. Holland, who formalized its operation in the volume Adaptation in Natural and Artificial Systems (MIT Press, 1992 – first edition 1975). [90]

GAs are based on the principle of natural selection: a population of candidate solutions evolves over time, and the most "suitable" solutions, i.e. those that best satisfy the objective function, are more likely to survive and transmit their characteristics. This process, inspired by evolutionary biology, allows to explore large and complex solution spaces in an efficient and adaptive way. [91, 92]

Unlike traditional optimization methods, which operate on a single solution, GAs work on entire populations, improving them through selection, crossover, and mutation operators. Each solution is represented by a chromosome (a string of variables or "genes") and evaluated by a fitness function that measures how close the solution is to the goal. The best ones are selected and combined to create new individuals, who progressively replace the less promising ones. [91–93]

As described by Sastry and Goldberg (2005) [92], a typical genetic algorithm cycle includes seven main steps:

- *Initialization* of the population, often random but sometimes based on prior knowledge;
- *Evaluation* of each individual's fitness;
- *Selection* of the best solutions based on fitness (through roulette, tournament, ranking, etc.);
- *Recombination* (crossover) between chromosomes of two or more parents;
- Random *mutation* of one or more genes to maintain variety;
- *Replacement*, in which the new generation replaces the previous one;
- *Iteration* until a stop criterion is met (number of generations, fitness threshold, or stability).

According to Goldberg [92], the real strength of GAs lies in the synergistic combination of these operators: each, taken individually, is ineffective, but together they produce a system capable of generating innovation and continuous improvement.

The fitness function is the heart of the evolutionary process: it guides the selection and allows the algorithm to adapt to the problem progressively. This function can be derived from a mathematical model, a numerical simulation, or a composite index that measures the performance of a solution.

In recent decades, numerous authors have expanded and refined the genetic paradigm, developing more robust variants or hybrid versions that integrate learning logics or multi-objective optimization methods. [94–96]

Among the main advantages of GAs are their flexibility, global exploration capability, and robustness against non-linear or noisy problems. However, they also present known criticalities: the high computational cost, the sensitivity to parameters and the need for effective coding of the solution. [90–97]

Today, genetic algorithms find application in a wide range of fields: from engineering to economics, from bioinformatics to artificial intelligence. Their modular structure, ease of hybridization, and ability to learn from successive generations still make them one of the most versatile and effective tools for solving complex optimization problems. [90–97]

2.5.2. Integrated Approach Based on a Custom Optimization Algorithm and a Machine Learning Model

Following the typical structure of genetic algorithms, a customized optimization procedure was developed, designed to identify an optimal configuration of the pressure sensors in the water network. In this formulation, the population is composed of several sets of candidate sensors, each representing a possible combination of monitoring nodes. Each individual in the population corresponds to a set of sensors, while each gene represents a single sensor. The genetic structure of the algorithm therefore mirrors the physical configuration of the monitoring system.

The algorithm initiates the process by generating an initial population of random or hydraulically and topologically constrained configurations (e.g., avoiding redundant nodes). With each generation, individuals are evaluated through a fitness function, which measures the ability of each configuration to ensure accurate leak location.

The basic dataset used for the training and evaluation of the model was generated using the EPANET software and its extension in Python EPyT (see sections 2.3.1. and 2.3.2), considering a set of distinct leak scenarios, each characterized by a single leak localized in a different node of the network. For each scenario, the pressure values at the network nodes during the simulation period were extracted.

To make the dataset more realistic and representative of field operating conditions, lognormal pressure noise was added to the simulated data, approximating the uncertainties and fluctuations typical of real measurements. The specific parameters used for noise generation and hydraulic simulation will be detailed in the relevant case study sections.

Since the simulated data inevitably exhibit variability and noise, a preliminary pre-processing step [98] of the pressure data was applied in order to improve the reliability of the machine learning model and the stability of the fitness assessment. This phase included:

- the subdivision of the dataset into a training set (80%) and a test set (20%) through stratified partitioning, to ensure a balanced representation of the different leak classes;
- standardization of the data, constraining each variable to have zero mean and unit standard deviation, according to the:

$$h_{st,i} = \frac{h_{act,i} - \mu_{h_i}}{\sigma_{h_i}} \tag{10}$$

  where $\mu_{h_i}$and σ_{h_i} represent the mean and standard deviation of non-standardized pressures ($h_{act,i}$), respectively;

- the application of Principal Component Analysis (PCA) [99] to reduce the dimensionality of the data and attenuate the residual noise, keeping the most significant components in terms of explained variance.

Once the pre-processing phase is complete, the data are provided to the Decision Tree model (for more details see section 2.4.1.), which is used as a learning and prediction tool as part of the evolutionary process.

For each individual generated by the genetic algorithm, the DT model is trained using the simulated pressure data relating only to the sensors belonging to the considered configuration, in order to learn the relationship between pressure variations and the location of the leak.

Once trained, the model is used in the predictive phase to determine the estimated location of the leak in the different simulated scenarios.

By comparing the model's predictions with the actual locations of the leaks, the localization accuracy associated with the analyzed sensory configuration can be calculated.

Finally, the average of the accuracies obtained on the different scenarios constitutes the performance indicator used for the evaluation of the individual's fitness.

The location accuracies obtained from the Decision Tree model are then weighted against the exposure levels of the different areas of the network, in order to introduce a risk-oriented optimization criterion.

In this context, the weight parameter, W_z, plays a central role from both an analytical and a decision-making perspective: it quantifies the relative importance of each zone z, allowing the model to prioritize areas where a failure to detect or an error in locating leaks would lead to more severe impacts (e.g., the formation of large sinkholes, structural damage, interruptions to the road network, etc.). Analytically, this parameter acts as an amplification factor for the leak localization accuracy in sensitive areas. Higher W_z values are assigned to the highest-risk zones in order to guide the algorithm toward a sensor configuration that ensures greater accuracy, and thus better monitoring, where the consequences of leaks on the urban fabric are estimated to be most critical.

It should be specified that in the first case study we will assume that the number of clusters (see equation 6) is equal to the number of junction nodes in which a leak can originate and therefore each single localization will refer to only one node.

On the basis of the above, the fitness function adopted (Weighted Accuracy, A_w) is as follows:

$$A_{\mathrm{w}} = \frac{\sum_z W_z \cdot \sum_{i \in z} A_{z,i}}{\sum_z W_z \cdot N_z} \tag{11}$$

where

- W_z represents the weight associated with the risk zone z;
- N_z is the number of leak scenarios in risk zone z;
- $A_{z,i}$ are the localization accuracies, obtained from the Decision Tree on the test set, for nodes belonging to risk class z.

For zoning with 3 levels of risk, z ∈ {R1, R2, R3}, we obtain:

$$A_w = \frac{W_{R1} \cdot \sum_{i \in R1} A_{R1,i} + W_{R2} \cdot \sum_{i \in R2} A_{R2,i} + W_{R3} \cdot \sum_{i \in R3} A_{R3,i}}{W_{R1} \cdot N_{R1} + W_{R2} \cdot N_{R2} + W_{R3} \cdot N_{R3}} \tag{12}$$

where:

- W_{R1}, W_{R2}, W_{R3} represent the weights associated with low, medium and high risk areas respectively. The selection of these values reflects the risk management strategy: for instance, assigning $W_{R3} > W_{R1}$ forces the genetic algorithm to favor sensor configurations that maximize accuracy in critical zones, even at the expense of lower accuracy in low-risk areas;

- N_{R1}, N_{R2}, N_{R3} indicate the number of nodes potentially subject to leak in the three risk classes;
- $A_{R1,i}, A_{R2,i}, A_{R3,i}$ are the localization accuracy values, obtained from the Decision Tree on the test set, for the nodes belonging to each risk class.

This formulation allows the performance of the sensor system to be comprehensively evaluated, rewarding configurations that guarantee greater localization accuracy in areas with higher risk. The genetic algorithm then proceeds iteratively by selecting, recombining and changing the best performing individuals, until it converges towards an optimal set of sensors capable of maximizing the A_w function (Figure 19).

The end result is a sensor configuration that not only optimizes the ability to detect leaks but does so in a way that is aware of territorial risk, offering an integrated and flexible approach to optimizing hydraulic monitoring.

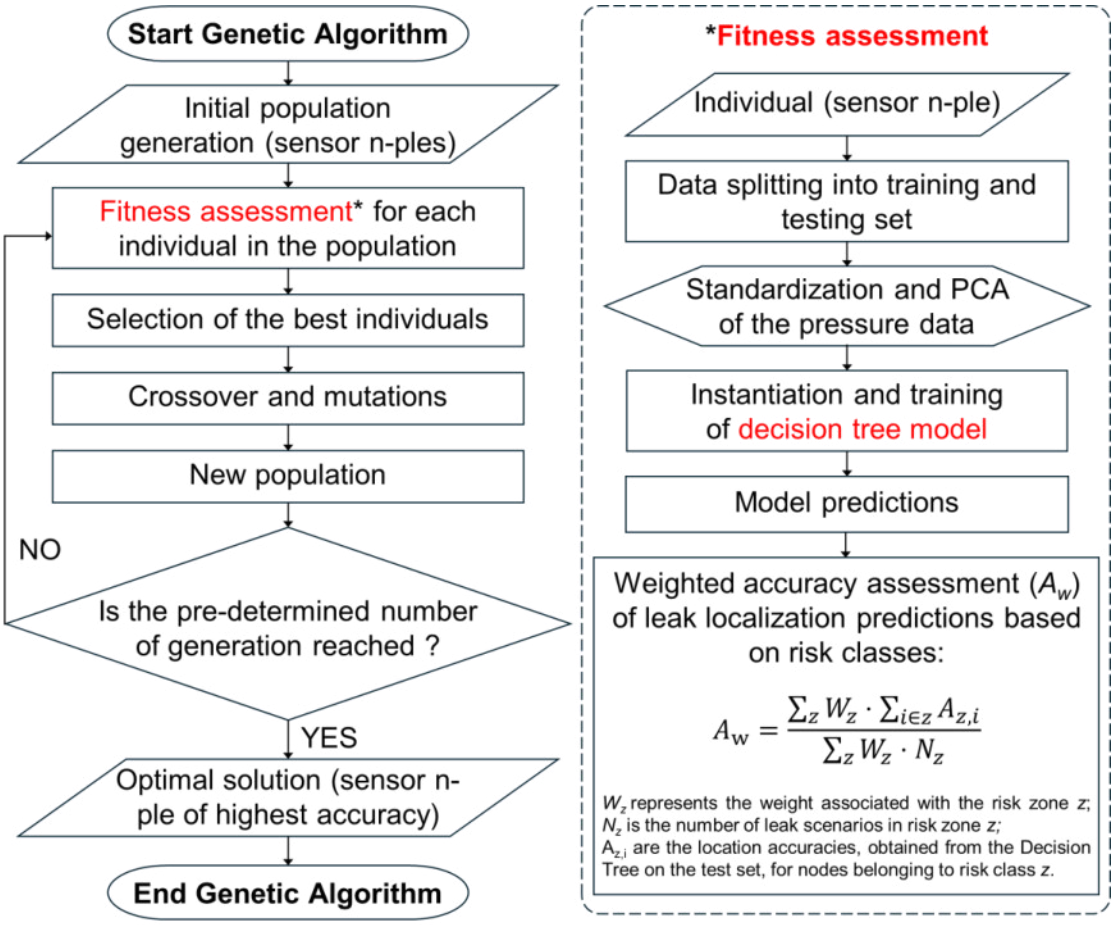

Figure 19. Logical scheme of the integrated approach based on customized genetic algorithm and Decision Tree (DT) model for the optimization of sensor placement based on HDL risk zoning. Adapted from Medio et al. (2024) [35].

2.5.3. Integrated Approach Based on the Pymoo Optimization Algorithm and a Model-Based Method

The second approach proposed differs from the previous one mainly for the model-based nature of the localization method and for the use of the evolutionary framework Pymoo [89] as an optimization tool.

Pymoo is an open-source framework developed and maintained by Julian Blank, a researcher affiliated with Michigan State University's Computational Optimization and Innovation Laboratory (COIN), under the supervision of Kalyanmoy Deb. The framework is supported by the AnyOptimization research community and was created with academic purposes, in order to provide advanced tools for solving and analyzing multi-objective evolutionary optimization problems. Today, Pymoo is increasingly recognized in the scientific community for its stability, efficiency, and reproducibility, making it a reference tool for optimization applications in engineering and science.

While maintaining the same general evolutionary logic as the approach described in Section 2.5.2, in this case the evaluation of sensor configurations is carried out by means of a model-based approach, which replaces the machine learning algorithm seen above with an analysis based on the sensitivity matrix and cosine similarity between the theoretical pressure signatures and the observed residues.

In this phase, the reference dataset was generated using EPANET (section 2.3.1.) and the Python libraries EPyT and WNTR (sections 2.3.2 and 2.3.3). Compared to the first approach, the introduction of WNTR represents a significant novelty: thanks to its advanced simulation capabilities, it has been possible to model leaks in a more realistic and rigorous way, ensuring greater control over both the location of the leak along the pipelines and the temporal variability of the dispersed flow.

Again, the leak scenarios have been constructed by assuming a single active leak at a time, but the leak is now considered along the pipeline rather than at the nodes. The dataset used for the construction of the theoretical signatures is characterized by the location of the leak in the centerline of the pipeline, while in the test phases its position is delocalized along the same stretch.

This setting allows you to evaluate the model's ability to correctly identify the leak pipeline, regardless of the exact location of the fault point. Further details on hydraulic parameters and simulation conditions are reported in the relevant section dedicated to the case study.

The localization model is based on the construction of a sensitivity matrix, S, as described in subsection 2.4.2, which quantifies the pressure variation in each node of the network following a hypothetical leak on each pipeline.

As already seen above, generally, the sensitivity matrix is subsequently normalized and binarized to generate a signature matrix of losses, through the application of a threshold that distinguishes between significant and negligible effects. However, the choice of threshold value is critical: thresholds that are too low make the matrix excessively dense, while high thresholds make it almost diagonal, reducing the discriminating capacity of the model.

To overcome this limitation, a threshold-independent measurement based on cosine similarity (see equation 9, subsection 2.4.2.) was adopted in the present study, which evaluates the directional coherence between the pressure residue vector r and the sensitivity vectors s_N.

This indicator, varying between −1 and 1, measures the directional alignment between the observed and theoretical pressure variations, making it more robust to noise and measurement uncertainties. A crucial advantage of this formulation is that it does not require the definition of arbitrary thresholds, thus avoiding the need for further internal optimizations. This is particularly advantageous in the context of optimizing sensor placement, as it allows the performance of candidate configurations to be directly evaluated without introducing an additional layer of complexity. In this way, the evolutionary process focuses exclusively on the optimal arrangement of the sensors, significantly reducing the overall computational load.

A further innovation of this approach concerns the performance evaluation metric, which replaces simple classification accuracy with a continuous measurement of the minimum hydraulic distance ($D_{hydraulic}$) between the pipeline actually subject to leakage and the one on which the leak prediction was made.

The distance is calculated along the minimum hydraulic path in the network graph:

$$D_{hydraulic} = 0.5 \cdot L_{leak} + min\_path(N_{leak}, N_{pred}) + 0.5 \cdot L_{pred} \quad (13)$$

where:

- L_{leak} and L_{pred} represent the lengths of the actual and predicted pipes affected by leakage, respectively;
- N_{leak} and N_{pred} indicate, respectively, the pairs of end nodes of the pipes affected by the actual and predicted leaks. The code compares all four possible combinations between these nodes (start–start, start–end, end–start, and end–end) and selects the shortest path. In this way, the calculation effectively considers the distance between the closest points of the two pipes;
- *min_path* represents the distance along the minimum hydraulic path (weighted by the actual length of the pipes) in the network graph connecting the end nodes (N_{leak} and N_{pred}) of the two pipelines. The hydraulic graph is generated from the EPANET model through the WNTR library;
- The result is a continuous measurement, expressed in meters, which represents the actual distance between the actual loss and the estimated one in an interpretable way. This formulation, which is closer to the physical behavior of the system, makes it possible to quantify the localization error in terms of the actual hydraulic distance along the pipelines, thus providing a more realistic and physically consistent metric than a simple classificatory evaluation.

The inclusion of half a length for each of the two pipelines ($0.5 \cdot L_{leak}$ and $0.5 \cdot L_{pred}$) in the calculation of the hydraulic distance it is justified by the fact that both the actual and the estimated leak are represented as leaks located in the central part of the pipes. Especially:

- The leak simulations used for constructing the theoretical signatures are carried out by positioning the leak at the midpoint of the pipe, corresponding to the central point of the hydraulic segment;
- The prediction of the leaking pipeline is expressed as a pipe identifier, even if the testing phase is conducted on a dataset with leaks off-center. This is because the objective is to identify the pipeline potentially affected by the leak, and not the exact location of the failure point along it. Hence the convention of considering half the length of the pipeline where the leak is expected when calculating the minimum distance, so as to always obtain representative and consistent distance values.

The hydraulic distances calculated for each leak scenario are then aggregated by zone as a function of the position of the pipeline that could be affected by a leak, in order to construct a weighted objective function that takes into account the importance of the different areas of the network.

Regarding the meaning and function of the weights W_z, reference is made to the previous subsection for a more detailed discussion. In this section, these coefficients are combined with the accuracy calculated in terms of minimum hydraulic distance D_z. Consequently, during the optimization process, the algorithm will assign greater weight to localization errors (i.e., the distances, $D_{z,i}$ between the actual and estimated leaks) in zones z where the risk is higher, in order to ensure more rigorous monitoring in the most critical areas.

Figure 20 reports the logical scheme adopted in the second case study, integrating the Pymoo optimization algorithm with a model-based method (sensitivity matrix and cosine similarity) to optimize sensor positioning as a function of risk and minimum hydraulic path. It is worth noting the overall similarity to the scheme adopted in the first case study, despite the use and replacement of new modules. This highlights the high degree of modularity and adaptability of the methodology across different problem settings.

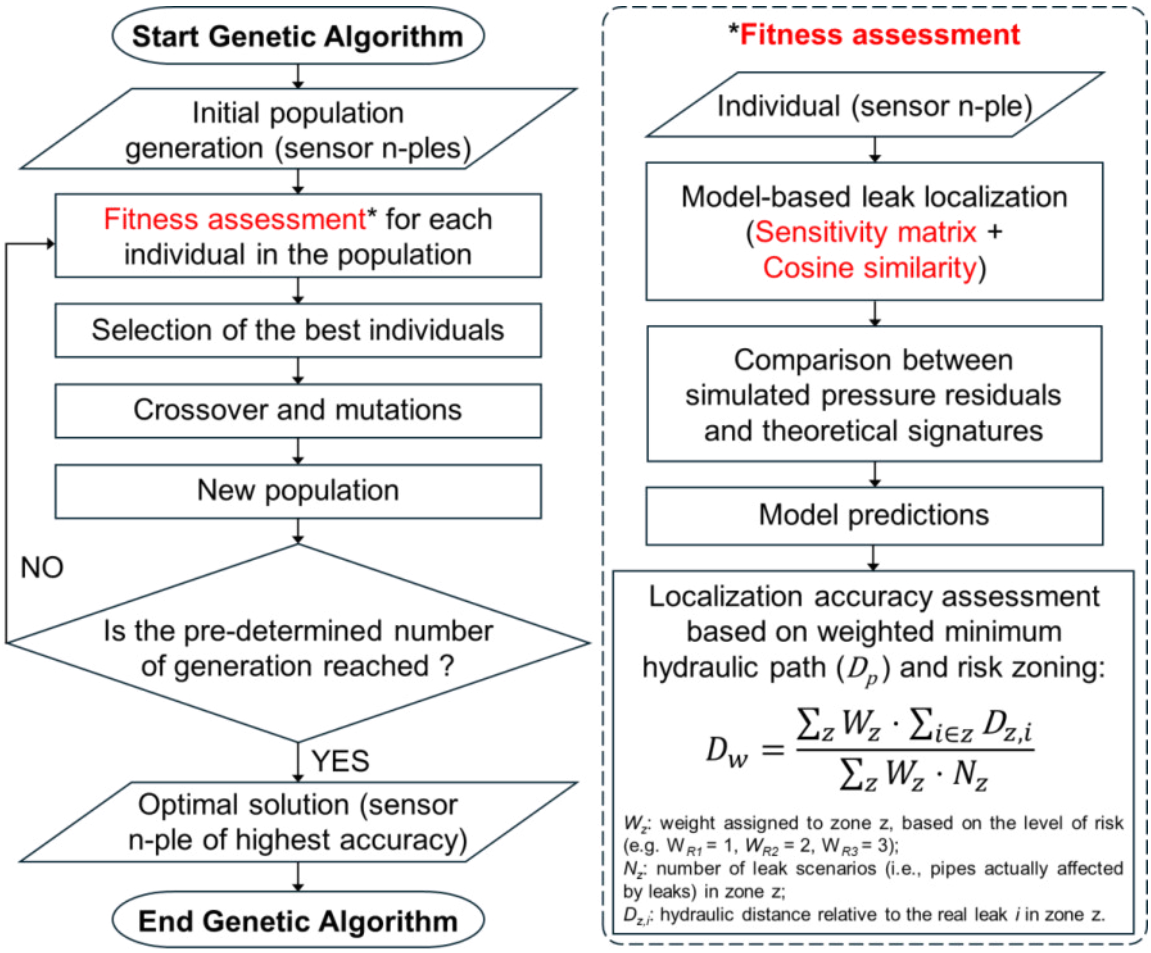

Figure 20. Logical scheme of the integrated approach based on Pymoo optimization algorithm and model-based model (sensitivity matrix and cosine similarity) for the optimization of sensor positioning as a function of risk and minimum hydraulic path.

In line with the risk zoning approach, the objective function to be minimized this time is defined as the weighted average of the minimum hydraulic distances D_w, in order to favor more precise solutions in the most critical areas:

$$D_w = \frac{\sum_z W_z \cdot \sum_{i \in z} D_{z,i}}{\sum_z W_z \cdot N_z} \tag{14}$$

where:

- W_z represents the weight associated with the risk zone z;
- N_z the number of leak scenarios in the risk zone z;
- $D_{z,i}$ the minimum hydraulic distance for the scenario i in the risk zone z.

For zoning with 3 levels of risk, z ∈ {R1,R2,R3}, we obtain:

$$D_w = \frac{W_{R1} \cdot \sum_{i \in R1} D_{R1,i} + W_{R2} \cdot \sum_{i \in R2} D_{R2,i} + W_{R3} \cdot \sum_{i \in R3} D_{R3,i}}{W_{R1} \cdot N_{R1} + W_{R2} \cdot N_{R2} + W_{R3} \cdot N_{R3}} \quad (15)$$

3. Case studies

In this thesis, two case studies have been developed, which will be presented in chronological order: first the *Real network 1* and then *L-Town*.

On both, a zoning of the Hydrogeological Disruption due to Leakage (HDL) risk was carried out and a subsequent optimization of the positioning of the sensors (OSP), aimed at maximizing the accuracy of leak localization in the areas of highest risk, with the ultimate goal of mitigating their negative effects on the urban fabric and the community.

The first case study, *Real network 1*, employs a custom genetic algorithm (GA) and a machine learning model for leak localization. Although it presented some modules of limited complexity, it allowed to build and validate the entire methodological process, constituting the experimental basis for the setting of the overall framework.

The second case study, *L-Town*, takes up the same conceptual framework and the same scientific objective, but introduces more advanced and rigorous modules, as a result of greater knowledge and awareness acquired. Especially:

- The GA custom is replaced by an evolutionary algorithm of literature, implemented through Pymoo;
- The leak location module is based on a model-based approach, based on a sensitivity matrix and cosine similarity;
- The evaluation metric has been further developed, considering the minimum hydraulic distance between the expected and actual leak. In this way, the performance of the model is evaluated continuously and realistically, surpassing the previous classificatory setting.

Overall, the presentation of both case studies allows to obtain a more robust validation and generalization of the proposed framework, while highlighting its modular nature and remarkable adaptability to different network configurations and operating conditions.

For the case studies under consideration, the following structure will follow:

1. Introduction and characteristics of the water distribution network;
2. Risk Zoning from HDL;
3. Modeling the hydraulic problem and generating pressure data;
4. Characteristics and parametric values of the proposed framework;
5. Results and discussion.

3.1. Real network 1

This section is partly based on the paper Medio et al. (2024) [35], co-authored by the PhD candidate.

3.1.1. Introduction and characteristics of Real Neetwork 1

The first case study analyzed concerns a real water distribution network, indicated in this work as *Real Network 1*.

This name is of a conventional nature, since, for reasons of confidentiality and protection of sensitive data, it is not possible to disclose the actual name of the municipality and the place where the network is located.

The network belongs to a medium-sized municipality in southern Italy, with an area of about 5 km² and a population of about 35 000 inhabitants, for an average density of about 6 600 inhabitants/km². In recent decades, there has been a steady population growth accompanied by intense building expansion. This phenomenon has led to the progressive disappearance of the original agricultural areas and a high level of soil sealing, with a consequent increase in the hydraulic and hydrogeological vulnerability of the territory.

The study area is characterized by a predominantly flat terrain, with altitudes between 100 and 150 m a.s.l., and weak slopes oriented towards the south.

From a morphological and geological point of view, the area is characterized by a stratigraphy dominated by pyroclastic deposits. The lithological succession includes layers of humified pozzolan, compact yellow tuff, breccias, lava fragments, xenoliths and cinerites, superimposed on more recent layers of ash, polychrome granular pumice, lapilli and volcanic sands. These materials, of an inconsistent nature and with high porosity, make the soil highly permeable but mechanically vulnerable, especially in the presence of load variations and prolonged infiltrations.

The local hydrographic system is mainly made up of artificial reclamation canals, built over the centuries to regulate the outflow of rainwater and encourage the cultivation of flat land. Among these, a single significant watercourse crosses the municipal area in question with a predominantly north-western trend, acting as the main collector of the surface water network. The configuration of the hydrographic system testifies to the agricultural and reclamation history of the territory, once characterized by a marked agricultural vocation, today largely compromised due to disorderly urban growth and the consequent reduction of permeable surfaces.

The climate of the area under consideration is Mediterranean, with long and hot summers and short and relatively mild winters. The rainfall trend is characterized by intense and concentrated rainfall between October and February, mainly due to the influx of humid air masses from the Tyrrhenian Sea. The average number of rainy days is about 87 per year, while the average annual rainfall is around 855 mm, with monthly values falling below 50 mm during the May-August period, sometimes close to zero in some years. Average winter temperatures are around 10 °C, while summer temperatures average 26 °C, with highs between 30 and 32 °C.

It is, therefore, a climatic context characterized by a marked thermal and rainfall seasonality, which affects both the water balance of the territory and the hydraulic behavior of the network.

The territory is now highly urbanized, with an almost continuous building fabric and a high population density. Residual natural areas and urban green spaces are limited and fragmented, while agricultural areas are now reduced to marginal portions. This context, together with the high degree of sealing and the reduced drainage capacity of the soil, contributes to increasing hydraulic vulnerability and making the management of surface and groundwater more complex.

The municipality's water distribution network has a total extension of about 23 km. The pipelines are mainly made of cast iron (74.9%) and grey cast iron (17.3%), with lower percentages of iron (4.3%), steel (2.8%) and polyethylene (0.7%). The diameters (Figure 21) vary between 53.6 mm and 406.4 mm, while the hydraulic model of the network includes 206 junction nodes, 231 pipelines and 7 injection points with almost constant piezometric load. From a topological point of view, the network has an almost symmetrical morphology, with regular rings distributed around the urban center.

Figure 21. Diagram of the Real Network 1 water network. The diameter classes of the pipes are represented by different colors, while the nodes and entry points are indicated by blue circles and green squares respectively. Reproduced from Medio et al. (2024). [35]

3.1.2. Real Network 1 HDL Risk Zoning

In this case study, HDL risk zoning has been carried out for mainly demonstration and methodological purposes, with the aim of illustrating the process of building risk maps consistent with the structure of the real water network.

A more complete and detailed procedure will be presented in the second case study.

Due to the limited availability of data relating to the hazard (H) and vulnerability (V) components, zoning was conducted considering only the exposure component (E).

However, it should be emphasized that, even if limited to the exposure component alone, this analysis should not be considered excessively simplistic. In fact, exposure represents the most immediately quantifiable component of risk, as it describes the socio-economic and infrastructural value of the areas and elements potentially affected by instability phenomena, measuring the extent of the potential damage according to the presence and value of the exposed assets, in particular those near the water network. On the other hand, the components of hazard (H) and vulnerability (V) are often more complex to assess, as they depend on less evident and not directly observable factors.

In the event that the components $H(x)$ e $V(x)$ are difficult to quantify, it is possible, as a precautionary measure as reported in paragraph 2.2, to adopt uniform measures throughout the territory in question.

In that case, they can be represented as constants, H_0 and V_0, getting:

$$R(x) = (H_0 \cdot V_0)\, E(x) = c_1\, E(x) \tag{16}$$

where c_1 is a positive constant that represents the uniform combination of hazards and vulnerabilities.

Under this hypothesis, risk and exposure are proportional point by point and share the same form of spatial distribution. Formally, for two generic points x_1 and x_2 of the domain, the relationship applies:

$$R(x_1) > R(x_2) \iff E(x_1) > E(x_2), \nabla R(x) = c_1\, \nabla E(x) \tag{17}$$

which expresses the preservation of the order and spatial gradient between the two variables.

It follows that the maps of R and E are topologically equivalent, i.e. they have the same trend and the same spatial variability, differing only by a constant scale factor.

If, in addition, the parameters H_0 and V_0 are normalized to the unit ($H_0 = V_0 = 1$), the constant c_1 reduces to 1 and the relationship takes the simplified form:

$$R(x) = E(x) \tag{18}$$

In this condition, the spatial distribution of risk coincides perfectly with that of exposure: the most exposed areas correspond directly to those at higher risk.

This formulation, while representing a simplification, allows to isolate the territorial effect of exposure as the main determinant of the spatial variability of risk, keeping the model interpretable and consistent with the nature of the available data.

The assessment of territorial exposure was conducted through the integration of demographic, cartographic and territorial data from official sources, flanked by satellite remote sensing products. This combination of sources allowed to build a coherent and up-to-date picture of land use and population distribution, subsequently validated through direct observations on high-resolution images:

- *Demographic and spatial data:* Used for estimation of population density and population distribution by census district;

- *Territorial information system:* used to enrich and detail the territorial analysis with additional information relating to buildings, road infrastructures, production areas, environmental sites of interest, etc. These data, integrated into the GIS environment through the WMS and WFS services, have ensured spatial coherence and local updating with respect to the ISTAT bases;
- *Satellite observations:* Used to verify and correct any recent discrepancies or changes in land use compared to official datasets.

Demographic and spatial data

The demographic information comes from the 2021 Permanent Census of Population and Housing carried out by ISTAT (Istituto Nazionale di Statistica) [100], which is an Italian public research body that deals with general censuses of population, services and industry, agriculture, sample surveys on households and general economic surveys at national level. Demographic information is disseminated at the sub-municipal level (census districts, sub-municipal areas, localities). The 2021 Territorial Bases (BT2021) update and complement the 1991, 2001 and 2011 editions. The data are mosaic at national level and released for individual regions.

The datasets are distributed in shapefile and (for 2011 and 2021) also in KMZ, with projection WGS 84 / UTM zone 32N (EPSG: 32632); attributes are UTF-8. The restitution scale varies indicatively from 1:5,000 (urban) to 1:25,000 (low-density areas). The data are fully usable with open-source GIS software (e.g. QGIS) and through the WebGIS IstatViewer (gisportal).

The BT2021 inherits and refines the geometric rules of the BT2011, with greater internal homogeneity and the introduction of encodings (e.g. Cod_Tipo_S) useful for reading land use/cover. The time reference is 31 December 2021; the administrative names in force on that date are documented by ISTAT.

The 2021 census districts at the national level amount to 756,376 (they were 402,677 in 2011), with a marked increase both in inhabited/productive locations and in scattered houses, confirming the jump in territorial granularity.

At the sub-municipal level, the placement and connection between population and housing in the Permanent Census is done by connecting the Basic Register of Individuals (RBI - Registro di Base degli Individui) to the Basic Statistical Register of Places (RSBL - Registro Statistico di Base dei Luoghi), which ensures unique and consistent geocoding for statistical units.

Resident population density (PD_i) of the i-th census section was calculated as:

$$PD_i = \frac{Pop_i}{A_{surf,i}} \tag{19}$$

where Pop_i denotes the number of residents (resident population) and $A_{surf,i}$ represents the surface area of the i-th census section.

To provide a sense of the spatial detail considered, it should be noted that the census section is the minimum territorial unit of the ISTAT spatial hierarchy. These sections are designed to be internally homogeneous and are typically delimited by physical and recognizable boundaries, such as the road network, railways, or natural elements (e.g., rivers and canals). In urban environments, this partitioning often leads to sections that correspond to a single city block (isolato), generally covering

a surface area smaller than 0.05 km^2. This high resolution ensures that the parameter PD_i acts as a precise micro-territorial proxy for anthropogenic exposure, attributing greater weight to areas with high building density and population concentration.

Territorial information system

The Territorial information system (SIT) [101] of the metropolitan authority of reference constitutes the institutional portal for the management, analysis and dissemination of geographical and thematic data relating to the territory of competence.

Through the dedicated web platform, developed to promote open access and transparency of public data, the SIT provides useful tools for various areas, such as those related to territorial planning, environmental monitoring, infrastructure management and so on.

The system allows data consultation through an interactive cartographic interface and provides various services ensuring full interoperability with the main GIS software.

Among the main services available are:

- Interactive map: allows direct visualization of the geographical data on the appropriate platform;
- GeoCatalog search: Allows you to identify available thematic layers through keywords and access their metadata;
- Download layer: section dedicated to the distribution of spatial data in open format, intended for reuse and integration into GIS applications;
- WMS (Web Map Service): provides an HTTP interface for requesting and displaying georeferenced maps in the form of raster images;
- WCS (Web Coverage Service): allows the publication and download of continuous raster data, such as orthophotos and digital terrain models;
- WFS (Web Feature Service): allows the consultation, extraction and interchange of vector data, maintaining the geometric and attributive structure of the original themes.

The SIT therefore represents a source of information at local level capable of providing additional details of an infrastructural, environmental and urban planning nature.

In the present work, this information has been used to integrate the analysis of territorial exposure, in particular through the use of layers relating to residential buildings, transport infrastructures, production and industrial areas, sensitive environmental sites and other components of the territory potentially vulnerable in the event of instability due to water leak.

The use of SIT data and themes is subject to compliance with the Creative Commons CC BY-SA 4.0 license, which allows its use, modification and redistribution, provided that the source is correctly cited and the same license is maintained for any derivative processing. In this document, the SIT data have been used exclusively for the purposes of research and territorial analysis, in full compliance with the conditions of use provided for by the license.

Satellite observations

The Google Earth Pro and Google Maps platforms [102, 103] have been used as visual validation tools for the verification and updating of information derived from other official datasets. These services, based on high-resolution satellite images and periodically updated aerial photographs, make it possible to observe the territory in detail and to conduct direct photointerpretation analyses aimed at improving the quality and timeliness of spatial data.

In particular, satellite observations were used to:

- Verify the spatial consistency between the land use classes reported in the official datasets and the real territorial configuration observable in the most recent images;
- Update and correct information on local changes in land use, such as the expansion of urbanised areas, the construction of new infrastructure or the conversion of agricultural and productive areas;
- Support the visual photointerpretation of areas potentially subject to impacts in the event of hydrogeological instability or failures in water pipelines.

The combined use of Google Earth Pro (for three-dimensional and morphological evaluations) and Google Maps (for the planimetric consultation of updated orthoimages) has made it possible to integrate quantitative and qualitative data, improving the precision and reliability of territorial mapping.
Satellite observations have therefore provided an independent level of verification and updating, which is essential to ensure the consistency and up-to-date nature of the information framework used in the assessment of exposure.

Classification criterion

Exposure value measures the negative consequences that an adverse event (in our case, a hydrogeological instability caused by water leaks) can have on the community, infrastructure, and essential services. This value therefore depends not only on the presence of physical elements, but above all on their function, strategic importance, and level of crowding.

For the definition of a consistent classification criterion, reference was made to the NTC2018 (Norme Tecniche per le Costruzioni - Technical Standards for Construction), the fundamental regulatory text governing the design, execution, and testing of building and infrastructure works in Italy [104], which divide the works into four classes of use according to their strategic importance (Table 2.), the degree of crowding and the consequences of an interruption of operation or a possible collapse.

The criterion defined by the NTC 2018 was therefore considered adequate to represent, in a homogeneous way, the relative exhibition value of the anthropic and infrastructural components of the territory, while remaining adaptable to any alternative schemes (e.g., based on urban planning categories or economic indices).

Table 2. Classes of use according to NTC 2018

Classes of use	Description
Class I	Buildings with only occasional presence, agricultural buildings.
Class II	Buildings whose use involves normal crowding, without contents dangerous to the environment and without essential public and social functions. Industries with activities that are not hazardous

	to the environment. Bridges, infrastructural works, road networks not falling under Use Class III or Use Class IV, railway networks whose interruption does not cause emergency situations. Dams whose collapse does not cause significant consequences.
Class III	Buildings whose use involves significant crowding. Industries with activities that are hazardous to the environment. Extra-urban road networks not falling under Use Class IV. Bridges and railway networks whose interruption causes emergency situations. Dams relevant for the consequences of their eventual collapse.
Class IV	Buildings with important public or strategic functions, also with reference to the management of civil protection in the event of disasters. Industries with activities that are particularly dangerous for the environment. Type A or B road networks, referred to in Ministerial Decree no. 6792 of 5 November 2001, "Functional and geometric standards for the construction of roads", and type C when they belong to routes connecting provincial capitals not also served by type A or B roads. Bridges and railway networks of critical importance for the maintenance of communication routes, particularly after a seismic event. Dams connected to the operation of aqueducts and electricity production plants.

In the present study, classes III and IV have been merged into a single category, since in many urban or peri-urban areas highly strategic structures with important public functions are absent or not very widespread, generally occupying small surfaces compared to the overall built fabric.

This choice allows to simplify the representation of the exposure value without compromising the significance of the analysis, while maintaining the distinction between low, medium and high exposure areas.

The exposure classes adopted are therefore as follows:

- E1 – Low exposure: Class I areas or buildings (temporary, agricultural or sparsely urbanized use);
- E2 – Medium exposure: Class II areas or buildings (constructions or activities with normal crowding);
- E3 – High exposure: class III–IV areas or buildings (buildings with significant crowds, highly strategic structures or with important public functions).

This classification allows a differentiated level of exposure to be assigned to the elements potentially exposed to the risk under consideration, i.e. those located near the water network. This distinction is decisive for the correct zoning of the territory and for the subsequent assessment of the associated risk.

Buffer zone definition and exposure attribution

A key element of the methodology is the creation of a buffer zone along the pipeline network. The buffer zone represents the potential influence of the WDN, i.e., the area that can be directly affected by hydrogeological instability caused by leaks.

In the present study, the width of the buffer was not defined on the basis of the physical extent of the instability, which in most cases is spread over a few meters, except for exceptional events that can reach a few tens of meters, but on the functional repercussions that a leak event can generate on the surrounding context. In fact, it was considered that even limited damage can produce the interruption of the road network, as well as temporarily compromise the usability of adjacent infrastructures,

leading to the unusability of buildings, suspension of economic activities and the need for restoration and excavation interventions.

Therefore, the width of the buffer was defined with the aim of including the areas immediately adjacent to the pipeline, including both the roads along which the water networks normally develop, and the adjacent structures and activities. With the need to represent an adequate band of influence, a conventional value of 25 meters per side with respect to the axis of the pipeline was adopted. This extension makes it possible to realistically represent the area of possible territorial and functional impact, also taking into account the restoration operations.

GIS Procedure for Exposure Zoning

The zoning of the exposure was conducted using a GIS procedure divided into four main phases, aimed at combining territorial zoning and network geometry:

1. Buffer construction around each main pipe section of the water network.
2. Preliminary zoning of the municipality in which the network extends, according to the classification criterion into three exposure classes (E1–E3) described above, deriving from the integration between ISTAT 2021 Territorial Bases, Territorial information system (SIT) and satellite observations.
3. Spatial intersection between the exposure zoning and the pipeline buffers, in order to transfer the prevailing exposure level present in the crossed territory to the pipeline adjacent areas.
4. Inheritance of the exposure class to the junction nodes (and/or pipelines) of the network that fall within the relevant buffer band.

The result is a water network that is "weighted" according to the surrounding urban and infrastructural context, in which each section and each node has a localized exposure value.

Results and interpretation of the Real Network 1 zoning

Below is the exposure map obtained (Figure 22), which shows a clear topological coherence between the water network and the urban structure of the analyzed territory:

- The central sections of the network have exposure values of E3, inheriting the high population density, the presence of important and historic buildings and a road system characterized by high criticality. Maximum exposure is also found at intersections with transport infrastructures (e.g. railway networks) or near public and strategic structures.
- The sections that cross ordinary urbanized areas are classified as E2, reflecting an average level of anthropization.
- Pipelines that cross marginal or agricultural areas are associated with class E1, indicative of low exposure.

This configuration coherently reflects the distribution of the built environment and the population, highlighting the close correlation between urban density and exposure to HDL risk.

Although this is a simplified zoning, it constitutes the basis for the integration of the risk component in the subsequent processes of optimising the positioning of the sensors, allowing for the attribution of higher priority weights to the areas with greater exposure value.

In this way, exposure mapping represents the first step towards an integrated HDL risk assessment, in which the territorial distribution of potential impacts is directly linked to the physical structure of the water network.

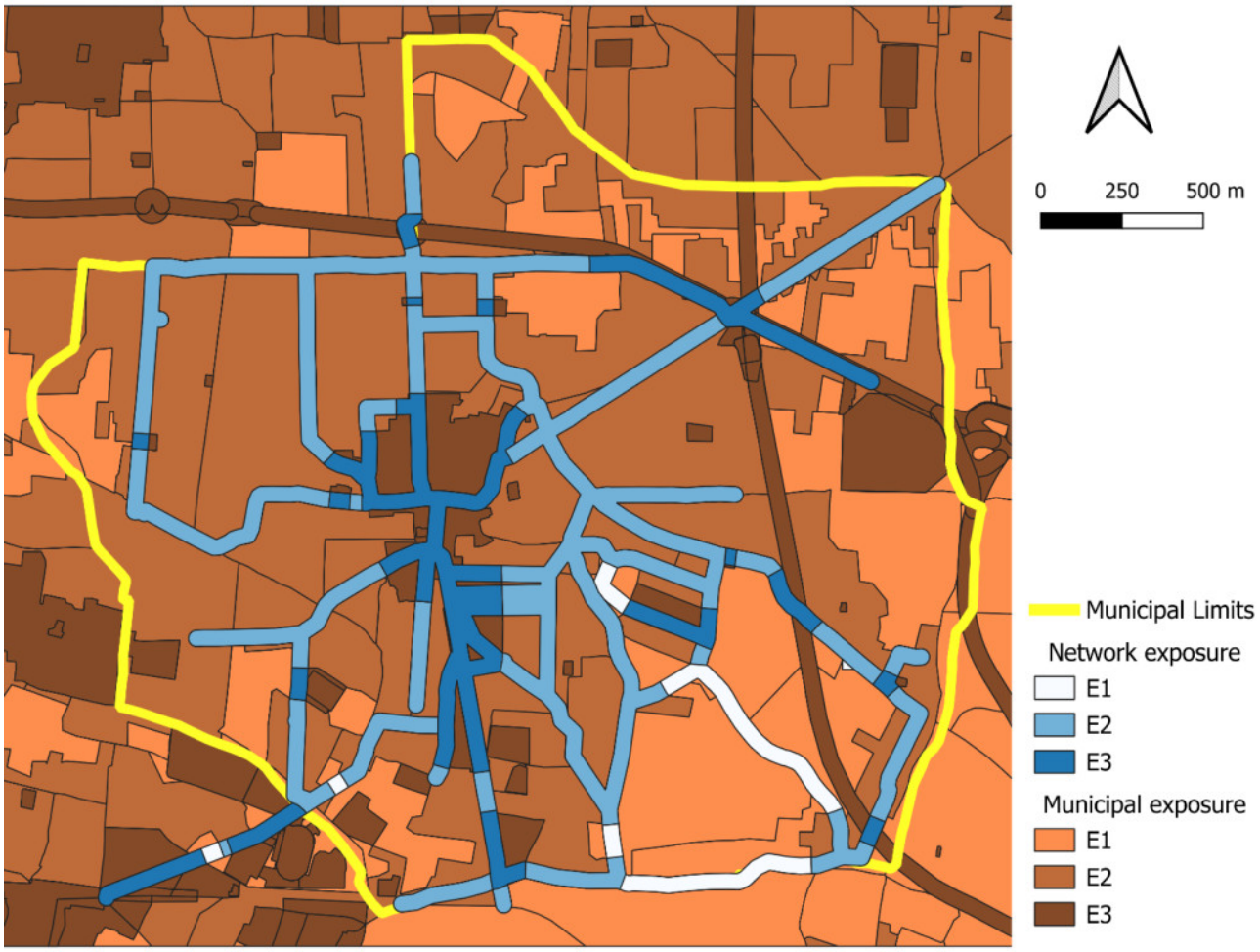

Figure 22. Zoning of territorial exposure (classes E1–E3) derived from the integration of ISTAT data, SIT and satellite observations. The pipelines inherit the prevailing exposure class of the area crossed, making it possible to distinguish the low (E1), medium (E2) and high exposure (E3) sections. The municipal perimeter is represented in yellow. Reproduced from Medio et al. (2024). [35]

3.1.3. Hydraulic modeling and pressure data generation for Real Network 1

The hydraulic modeling of *Real Network 1* was carried out using EPANET 2.2 and the EPyT library, already presented in paragraph 2.3..

These tools enabled the automatic execution of multiple simulations and the generation of synthetic pressure data for different leak scenarios.

Different hydraulic scenarios were generated, each corresponding to a specific leak condition.

The main operating assumptions adopted are:

- the leaks are located exclusively in the junction nodes;
- each scenario has only one active leak;
- the total number of scenarios coincides with the number of nodes in the network;
- generation of a reference scenario in the absence of leaks (No Leak);
- the duration of each simulation scenario is 50 days, with a constant time step.

The results of all the simulations were collected in a single synthetic dataset, which was subsequently used to train the machine learning model for the automatic localization of leaks.

Demand and leak modeling

The water demand at the nodes was represented by a variable daily pattern, applied uniformly to the entire network.

The demand $q_i(t)$ (l/s) at time t and at junction node i was calculated as the product of the demand coefficient $DC(t)$ and of the basic demand of the node q_{B_i} according to the relation:

$$q_i(t) = DC(t) \cdot q_{B_i} \tag{20}$$

The pattern (Figure 23) shows a typical urban trend, with morning peaks and night-time minimums, while the stochastic variability is described by a lognormal distribution with coefficient of variation, $CV = 0.2$.

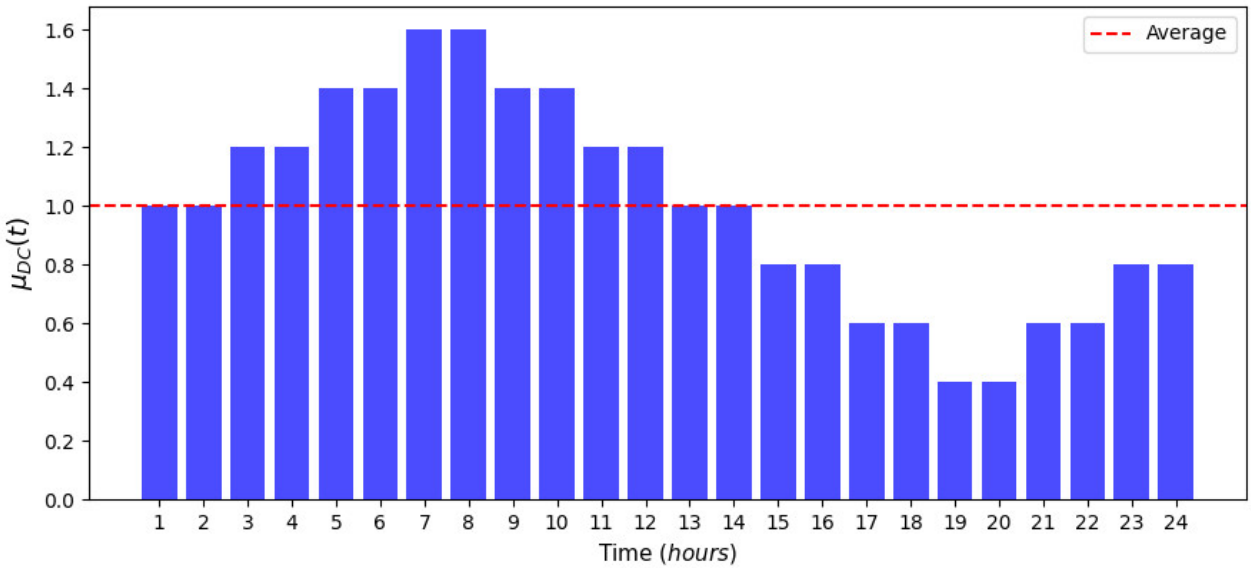

Figure 23. Daily trend of the average demand coefficient $\mu_{DC}(t)$. The red dotted line represents the daily average value. Reproduced from Medio et al. (2024). [35]

The leaks were simulated by emitters placed in the nodes, whose flow rate Q_i depends on the piezometric head h_i and the emission coefficient EC_i:

$$Q_i = EC_i \cdot h_i^{0.5} \tag{21}$$

The EC_i values were selected to realistically represent a localized leak in a pipeline; for further details, refer to [35]. Different values of the emitter coefficient can be found in the literature [66, 105].

Structure of the generated dataset

The final dataset (Table 3.) includes both the no-leak scenario (*No Leak*) and the various localized leak scenarios (*Leak Node i*). Each row represents a specific time point and contains the pressures at the main nodes of the network. The following table shows, for example, two consecutive time points for the *No Leak* scenario and two for the *Leak Node i* scenario.

Table 3. Leak scenario datasets – Real Network 1

Scenario Number	Scenario	Tempo [s]	Time Scaled [s]	Node Pressure 1 [m]	...	Node Pressure i [m]	...
0	No Leak	864000	0	78.655	...	82.171	...
0	No Leak	867600	3600	77.313	...	80.828	...
...	...	...	...	...	...	...	...
¡	Leak Node i	986400	122400	89.464	...	82.556	...
¡	Leak Node i	990000	126000	89.976	...	83.202	...
...	...	...	...	...	...	...	...

This structure allows you to consistently represent the temporal evolution of pressures for each scenario and to clearly distinguish between leaky and non-leaking conditions. The dataset thus constructed constitutes the database for training the machine learning model (Decision Tree).

3.1.4. Characteristics and parameter values of the proposed framework for Real Network 1

The optimization framework used for *Real Network 1* is based on a customized genetic algorithm, designed to identify the optimal configuration of pressure sensors capable of maximizing the accuracy of leak location in areas with higher HDL risk.

The algorithm integrates a supervised machine learning model, in particular a Decision Tree Classifier, used to train the sensor system in leak localization based on simulated pressure data.

The localization ability of the different sensory configurations is then quantified using the weighted accuracy metric (A_w), which constitutes the fitness function of the genetic algorithm.

The general logic of the framework, together with the theoretical foundations relating to the data pre-processing stages, Decision Tree, etc., was described in Chapter 2, to which reference should be made for further details.

In this section, on the other hand, the implementation and parametric aspects adopted for the application to *Real Network 1* are illustrated in detail.

The optimization process begins with the random generation of an initial population of sensor configurations, each consisting of a distinct set of measurement nodes. Each individual represents a possible combination of sensors and is encoded as a vector of unique indices, so as to avoid the presence of duplicates.

For each individual, the Decision Tree model is trained using pressure values from the simulated dataset, including multiple leak scenarios.

Prior to training, the data is standardized using the *StandardScaler* function of the scikit-learn library [79, 106] and then reduced in dimensionality through Principal Component Analysis (PCA), maintaining a number of components equal to the number of sensors considered. This step reduces information redundancy and improves the stability and reliability of the classification model.

The localization accuracy obtained by the model represents the fitness function associated with each individual and is further weighted according to the risk levels of the three areas of the network, so as to favor the most effective solutions in the most vulnerable areas.

The overall evaluation metric, called weighted localization accuracy (A_w), therefore constitutes the fitness function that drives the entire evolutionary process. Based on the assumptions and simplifications expressed by equations (16), (17) and (18) it is possible to rewrite the weighted accuracy (12) in the following way:

$$A_w = \frac{W_{E1} \cdot \sum_{i \in E1} A_{E1,i} + W_{E2} \cdot \sum_{i \in E2} A_{2,i} + W_{E3} \cdot \sum_{i \in E3} A_{E3,i}}{W_{E1} \cdot N_{E1} + W_{E2} \cdot N_{E2} + W_{E3} \cdot N_{E3}} \tag{22}$$

where:

- W_{E1}, W_{E2}, W_{E3} represent the weights associated with low, medium and high exposure areas respectively;
- N_{E1}, N_{E2}, N_{E3} indicate the number of nodes potentially subject to leak in the three exposure classes;
- $A_{E1,i}, A_{E2,i}, A_{E3,i}$ are the localization accuracy values, obtained from the Decision Tree on the test set, for the nodes belonging to each exposure class.

Once fitness has been assessed, the population is evolved through the classic genetic operations of selection, crossover, and mutation, with the aim of progressively improving overall performance. In particular:

- *Crossover*: implemented according to a *one-point* scheme, in which the genes (i.e. sensor nodes) of two individuals are combined to generate new candidates;
- *Mutation*: applied with a 5% probability, randomly replacing a sensor within the individual to maintain genetic diversity.
- *Elitism:* a quota equal to 10% of the best individuals of each generation are automatically retained in the next.
- *Uniqueness check:* in each generation any duplicate individuals (twins) are eliminated, ensuring that all sensor configurations are unique.

The evolution continues until the maximum number of generations is reached (500 in the case analyzed), since it has been observed that, with this threshold, the accuracy tends to stabilize by reaching a horizontal asymptote, indicative of the achievement of an optimal or at least sub-optimal solution (for an example, see Figure 24).

At the end of the evolutionary process, the framework returns the optimal (or sub-optimal) configuration of the sensors together with the corresponding localization accuracy.

The problem constraints and the main parameters used for the genetic algorithm configuration are summarized in Table 4.

Table 4. Optimization constraints and genetic algorithm (GA) parameters for Real Network 1. [35]

Description	Value
Number of genes (sensors)	3
Population size	60 individuals
Crossover probability	100 %
Mutation probability	5 %
Elite percentage	10 %
Number of generations	500

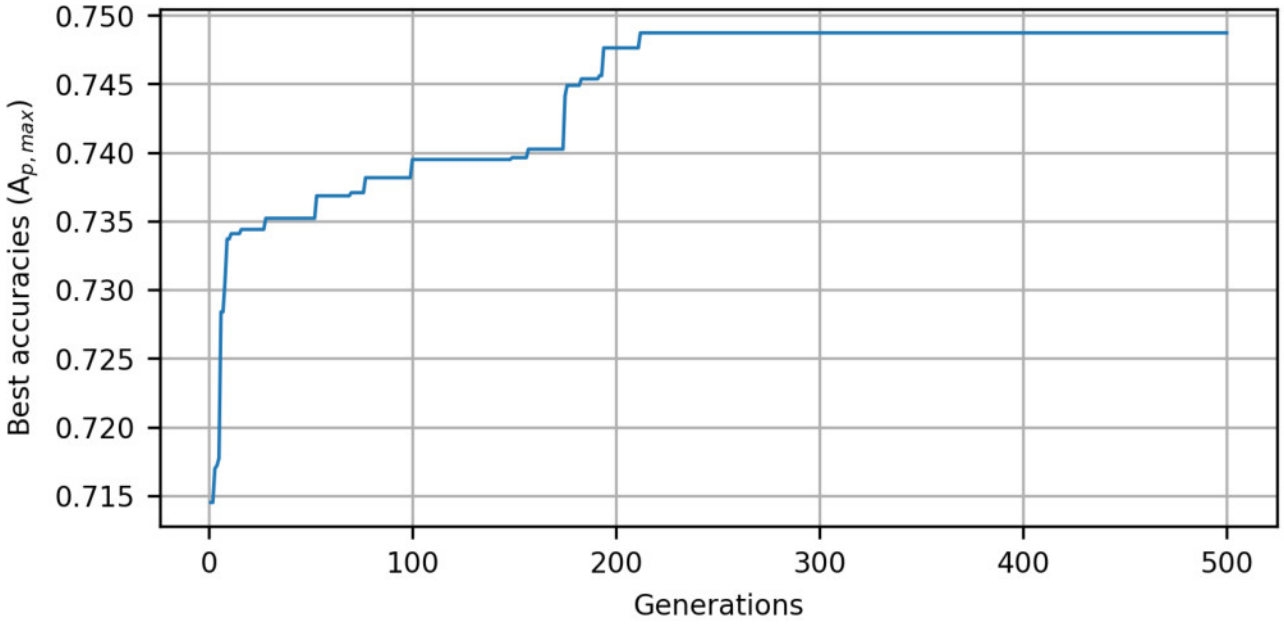

Figure 24. Trend of the maximum fitness function value (maximum weighted accuracy per generation $A_{p,max}$) during genetic evolution for the case without zoning (W_{E1} = 1 W_{E2} = 1 W_{E3} = 1). The progressive increase of the curve indicates the search for the optimal or sub-optimal solution for the positioning of the sensors. After about 200 generations, fitness tends to stabilize, suggesting the achievement of a configuration close to the overall optimum.

Overall, the proposed framework combines evolutionary research and machine learning into a single, coherent architecture that optimizes sensor placement while taking into account the areas at greatest risk from HDL.

3.1.5. Results and discussion for Real Network 1

It should also be remembered that the optimization of the positioning of the sensors was carried out taking into account the zoning of the HDL risk, through the assignment of weights representative of the different levels of exposure.

Since the model is based on this principle, different sets of weights have been tested in order to evaluate to what extent they influence the final distribution of the sensors and the overall performance of leak localization.

The choice of weights represents an aspect that is far from marginal: values that are too similar to each other tend to generate configurations similar to those obtained in the absence of zoning, while excessively high weights assigned to the highest risk areas, compared to those at lower risk, can determine an unbalanced concentration of sensors in the most exposed areas only, without bringing real benefits in terms of overall accuracy.

In this case, the adopted values were chosen because they allow for a sufficiently significant sensor deployment consistent with the spatial distribution of risk, while avoiding excessive imbalances. However, the purpose of this first application was not to precisely calibrate the optimal weights, but rather to demonstrate the general validity of the proposed method, postponing any more in-depth sensitivity analyses on the weights to future developments.

On the basis of these premises, the three exposure levels (W_{E1}, W_{E2}, W_{E3}) were assigned the values $(1,3,5)$, corresponding respectively to the areas of low, medium and high exposure to HDL risk.

The procedure described has been applied considering a number of sensors equal to three. The resulting optimal sensor configuration is shown in Figure 25.

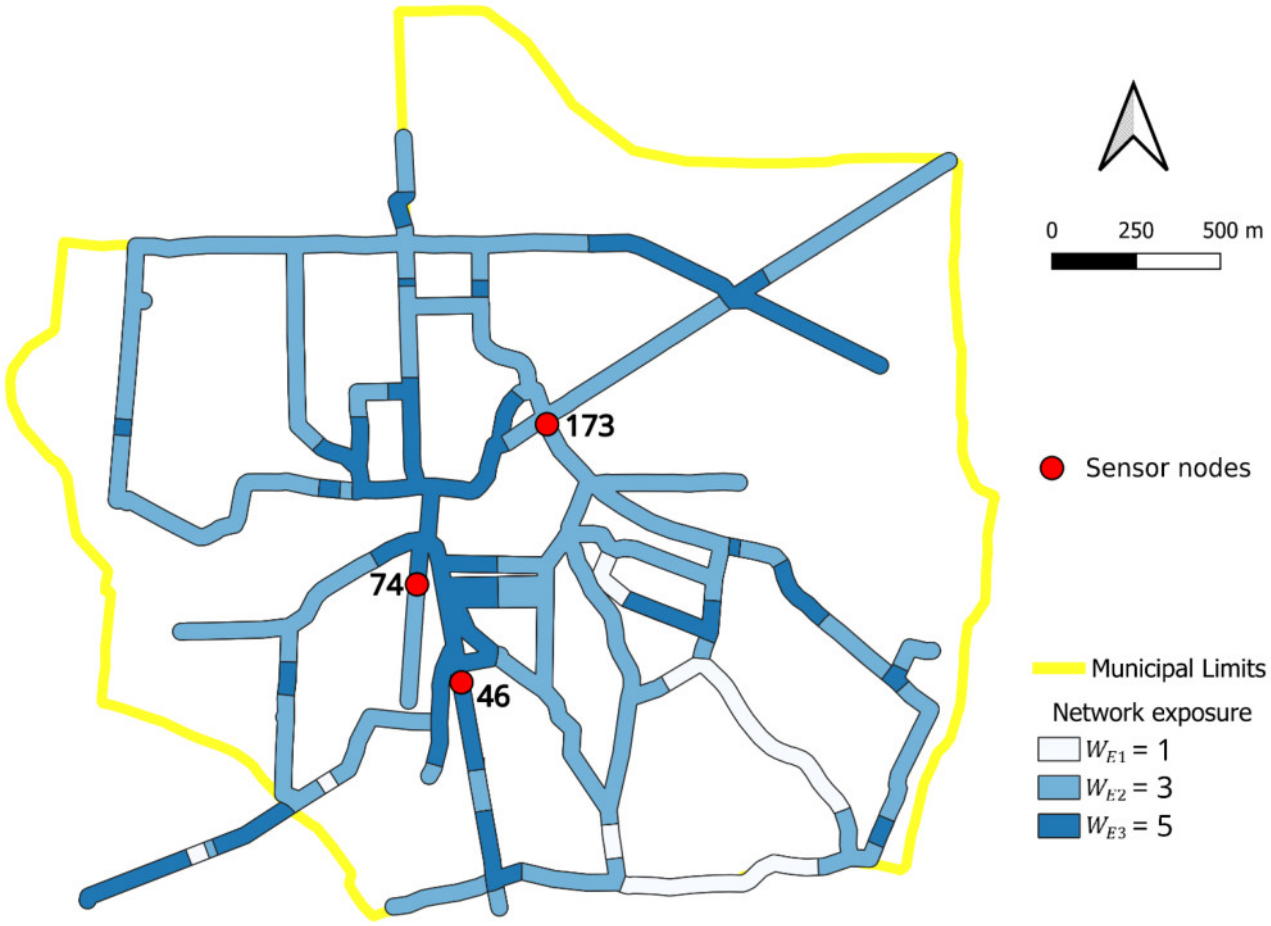

Figure 25. The map shows the optimal configuration of the sensor nodes (red circles) identified by the genetic algorithm, considering the subdivision of the network according to the real zoning of the exposure to HDL risk. The bands of influence are

colored according to the level of exposure ($W_{E1} = 1$, $W_{E2} = 3$ and $W_{E3} = 5$), with the highest values associated with the areas with the greatest exposure value. Adapted from Medio et al. (2024). [35]

The analysis shows that the global average localization accuracy (A_{GLOBAL}) on the entire network is equal to 0.740, while in the areas with greater exposure (class E3) it reaches 0.837.

To verify whether the algorithm is actually able to direct the optimization towards configurations that favor high-risk areas, increasing the probability of correctly identifying losses in these sectors, the same procedure was repeated assuming homogeneous exposure weights $(W_{E1}, W_{E2}, W_{E3}) = (1,1,1)$, equivalent to a uniform risk condition or, in other words, to a situation in the absence of risk differentiation.

In this scenario (Figure 26), the optimization goal becomes the maximization of the global average accuracy (A_{GLOBAL}) only, regardless of the location of the leak.

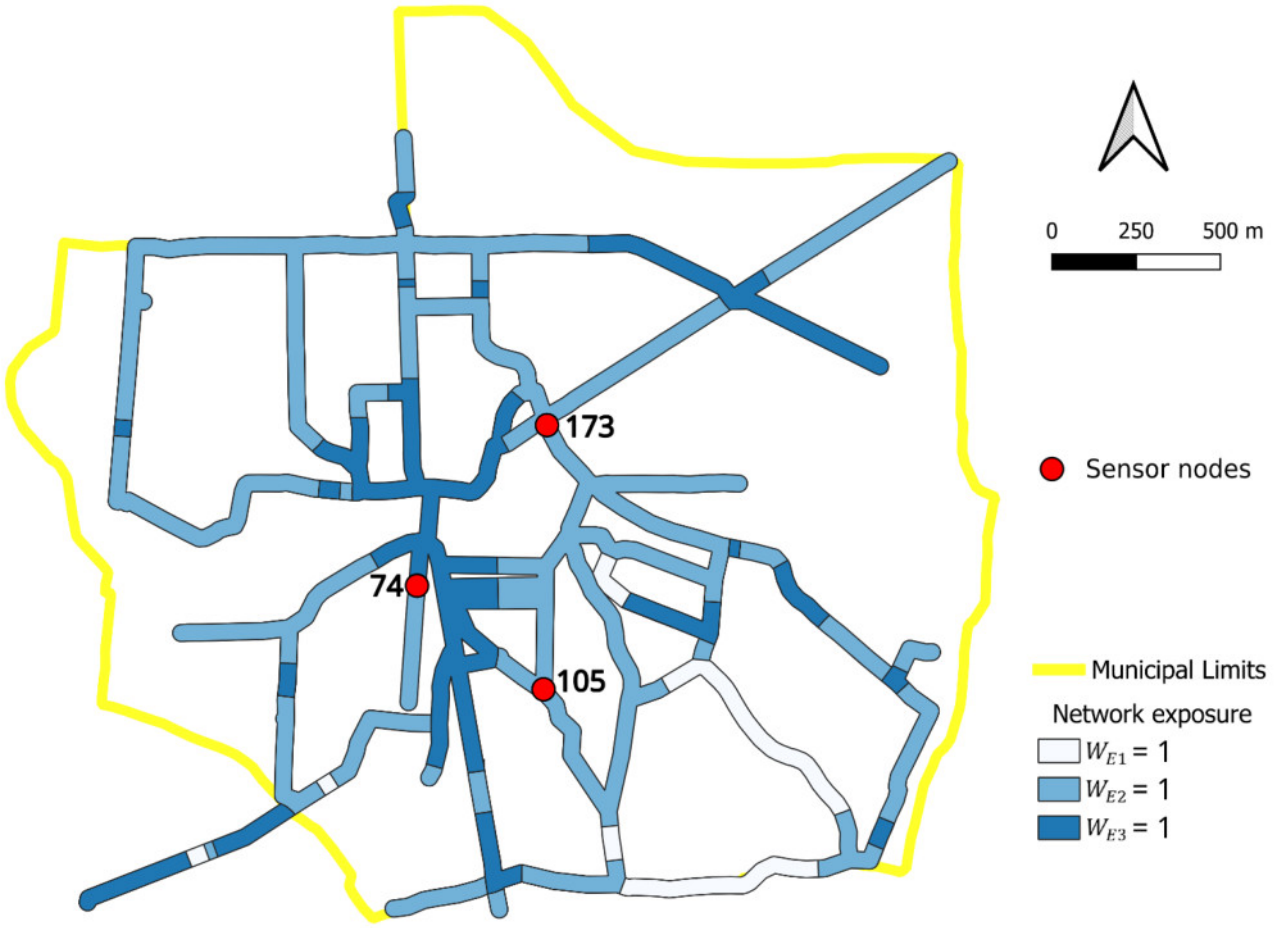

Figure 26. The map shows the optimal configuration of the sensor nodes (red circles) identified by the genetic algorithm considering all unit weights ($W_{E1} = 1$, $W_{E2} = 1$, $W_{E3} = 1$) which is equivalent to an optimization in the absence of zoning. Adapted from Medio et al. (2024). [35]

The results (Table 5) show a slight increase in global average accuracy (A_{GLOBAL} = 0.743), while the one referred to the E3 areas ($A_{E3,av}$) is reduced only marginally (from 0.837 to 0.829). Furthermore, the optimal positions of the sensors, illustrated in Figure 26, are almost concentrated in the central part of the network, similarly to the previous case, highlighting how the topological symmetry of *Real Network 1* tends to naturally constrain the optimal configurations. In this sense, risk weighing does not significantly change the overall performance but introduces greater environmental and territorial coherence in the distribution of sensors.

Table 5. Global and local localization accuracy per zone, with different weight values relative to the real zoning. [35]

W_{E1}, W_{E2}, W_{E3}	**1, 3, 5**	**1, 1, 1**
Optimal sensor set (Node IDs)	46, 74, 173	74, 105, 173
Average localization accuracy (A_{GLOBAL})	0.740	0.743
Localization accuracy in E2–E3, ($A_{E2,av}$, $A_{E3,av}$)	0.700 – 0.837	0.707–0.829

The actual zoning of the network presents a distribution that tends to be symmetrical with respect to the urban centre, which explains why variations in the weights do not produce large shifts in the position of the sensors.

To highlight the behavior of the method under different conditions and test its flexibility, a fictitious and highly asymmetrical zoning was introduced, represented in Figure 27.

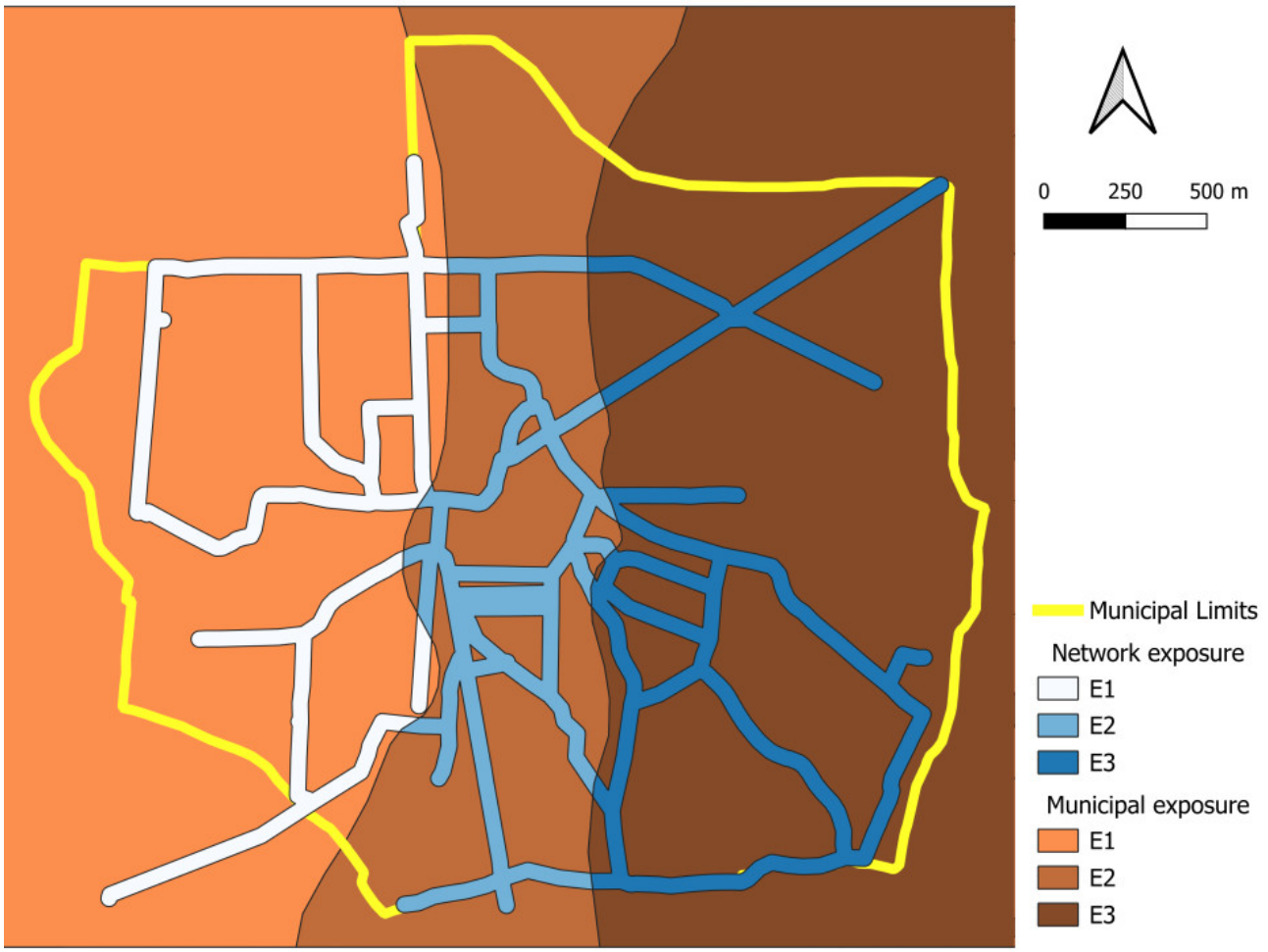

Figure 27. Fictitious zoning of Real Network 1. The zoning was created to obtain a clear asymmetry of the exposure values in order to more effectively test the proposed algorithm. Reproduced from Medio et al. (2024). [35]

In this configuration, the eastern part of the town has been defined as high exposure (E3), while the western part as low exposure (E1).

Also in this case, 3 sensors and weights $(W_{E1}, W_{E2}, W_{E3}) = (1,3,5)$ have been adopted, as shown below (Figure 28).

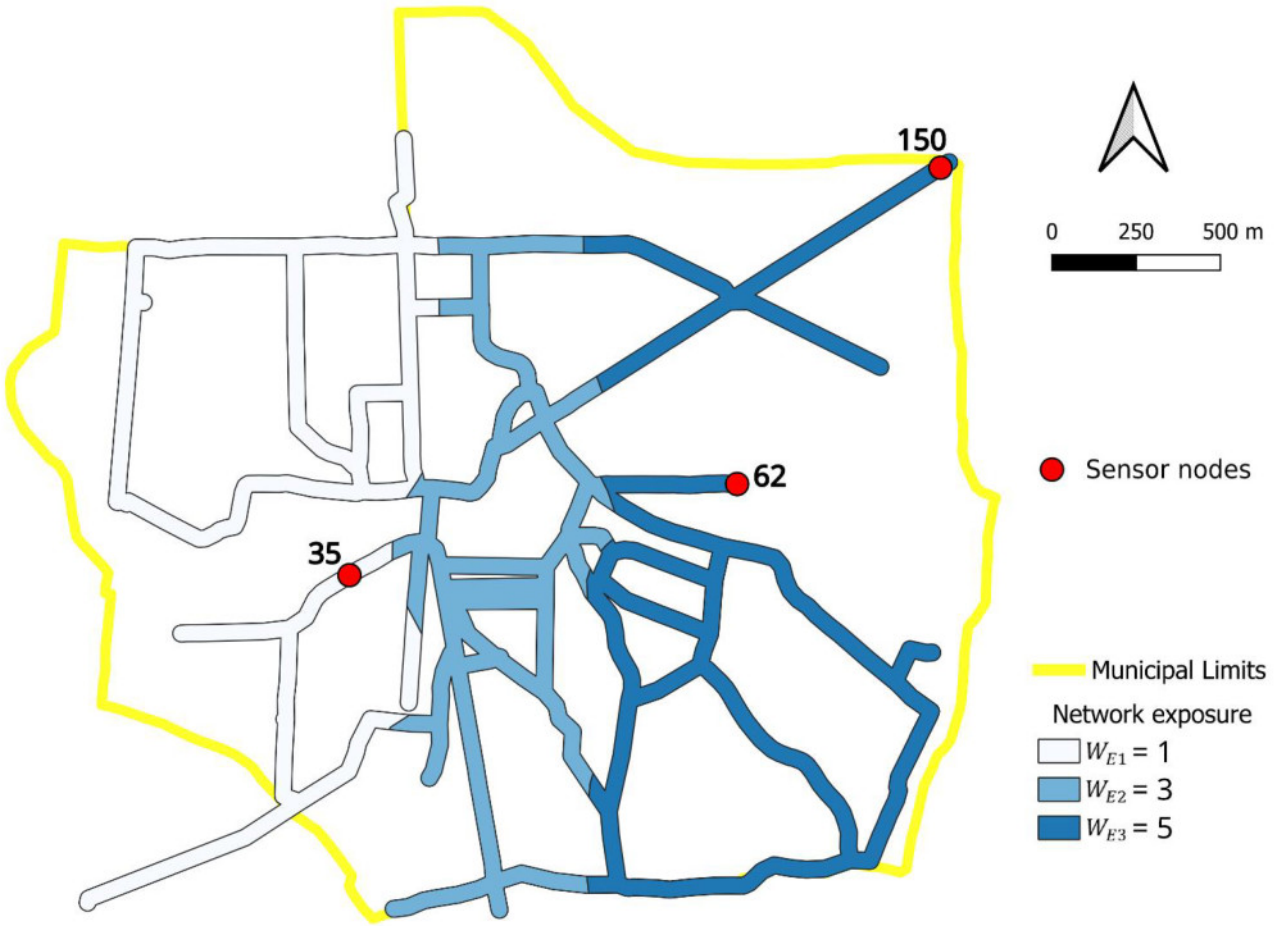

Figure 28. The map shows the optimal configuration of the sensor nodes (red circles) identified by the genetic algorithm, considering the subdivision of the network according to the fictitious zoning of the exposure to HDL risk. The bands of influence are colored according to the level of exposure (W_{E1}=1, W_{E2}=3 and W_{E3}=5), with the highest values associated with the areas with the greatest exposure value. Adapted from Medio et al. (2024). [35]

The results, shown in Table 6, for asymmetric zoning show a slight reduction in global average accuracy (A_{GLOBAL}) compared to that in the absence of zoning, confirming that the introduction of a high spatial asymmetry does not greatly compromise the overall ability of the method to effectively locate leaks; however, for certain areas there can be significant reductions in accuracy, as is the case for E2, which drops by 14 %.

Table 6. Global and local localization accuracy per zone, with different weight values relative to the fictitious zoning. [35]

W_{E1}, W_{E2}, W_{E3}	**1, 3, 5**	**1, 1, 1**
Optimal sensor set (Node IDs)	35, 62, 150	74, 105, 173
Average localization accuracy (A_{GLOBAL})	0.735	0.743
Localization accuracy in E2–E3 ($A_{E2,av}$, $A_{E3,av}$)	0.723 – 0.867	0.860 – 0.610

As shown in Figure 28, the asymmetry of the zoning leads two sensors (node IDs 62 and 150) to be located in the eastern high-risk zone (E3), while the third (node ID 35) is located in the western low-risk part (E1).

This configuration still guarantees good overall accuracy across the entire network (A_{GLOBAL}) but generates some extreme situations from the point of view of positioning: nodes with IDs 62 and 150

are in fact located at the end of two isolated branches, characterized by a high density of nodes. These traits tend to behave as poles of attraction for sensors, since their placement in those areas increases the average local accuracy in the E3 zone.

However, although this arrangement is numerically consistent, it is not necessarily the most appropriate solution in engineering terms. Therefore, especially in the presence of isolated sections full of knots, a targeted treatment aimed at limiting their influence in the optimization process would be appropriate.

For example, assuming that these traits are excluded as potential sensor locations, a more spatially balanced configuration is obtained (Figure 29), although accompanied by a decrease in numerical accuracy in E3 ($A_{E3,av}$ = 0.826). However, these aspects can be managed upstream through the use of more appropriate metrics and localization methods.

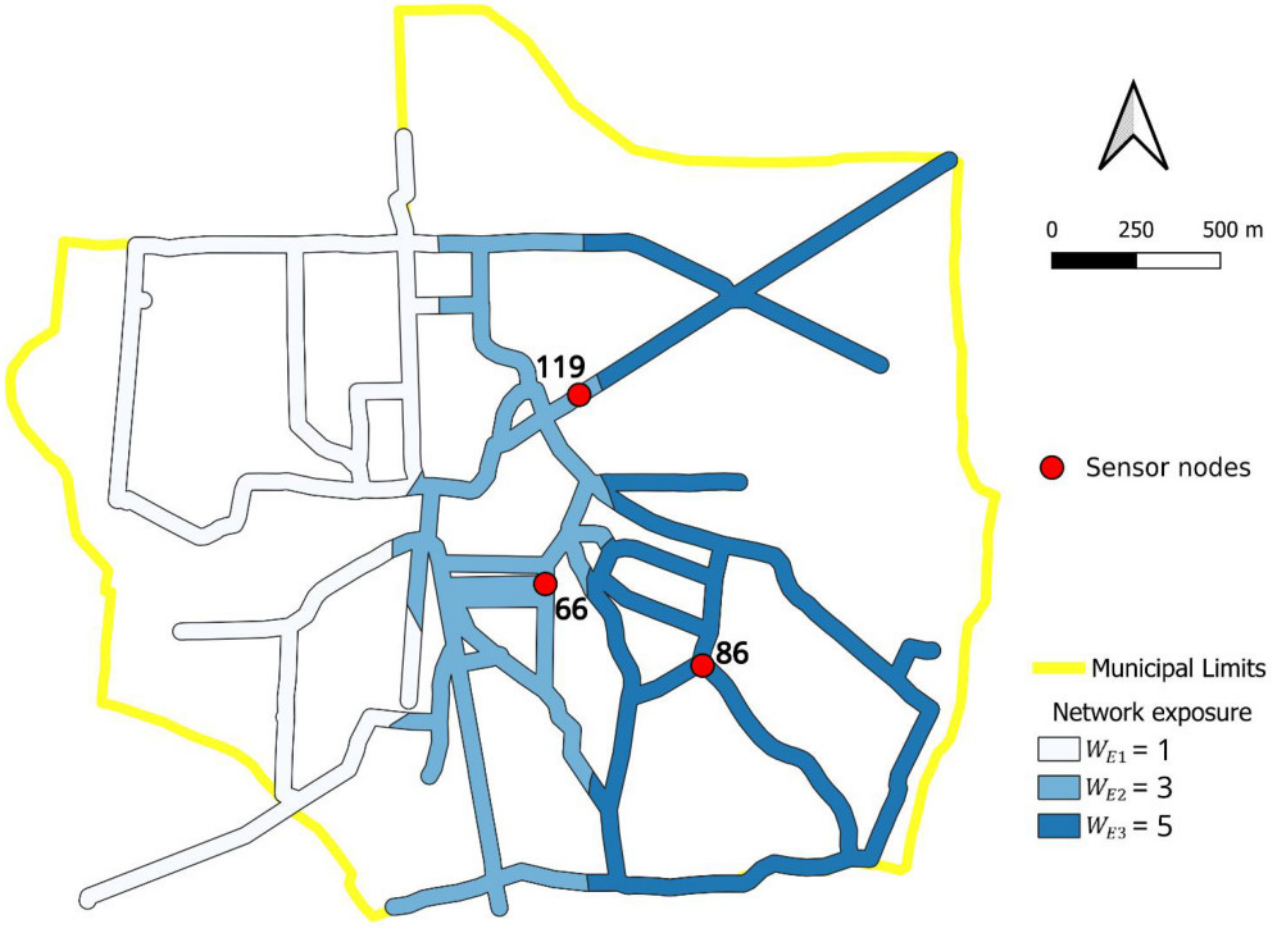

Figure 29. Optimal or (sub-optimal) sensor configuration for fictitious zoning (high asymmetry), in which isolated sections were excluded as potential sensor locations.

Therefore, here, the fundamental aspect to be highlighted is the ability of the proposed method to orient the positioning of the sensors on the basis of HDL risk zoning, thus expanding the potential of the framework in the mitigation of adverse phenomena related to water leaks.

Furthermore, the experiment demonstrates that the method does not distort the solutions obtained with a classical approach but integrates them with a component of territorial and environmental analysis.

Therefore, in cases where the actual zoning is symmetrical or evenly distributed, the differences compared to traditional configurations remain limited; conversely, in the presence of highly asymmetric zoning, the sensor placement may differ more markedly, allowing the positioning strategy to be adapted to specific local conditions.

In such contexts, a more targeted calibration of the weights could become a useful tool for finding the right system sensitivity to risk and further optimizing sensor distribution.

In summary, the results obtained for Real Network 1 confirm that the proposed framework effectively integrates the risk component into the optimisation process, ensuring a consistent and balanced distribution of sensors. The method allows for high overall localisation performance to be maintained, while offering greater attention to areas with higher levels of risk exposure and a more informed perspective on the urban environment and territorial risk.

3.2. L-Town

3.2.1. Introduction and characteristics of the L-Town network

The *L-Town* network represents one of the best known benchmarks for international research on the detection and location of water leaks. It was developed within the *BattLeDIM* (Battle of the Leakage Detection and Isolation Methods) competition, promoted in 2020 by the KIOS Research and Innovation Center of Excellence of the University of Cyprus (Vrachimis et al. 2022) [107], in collaboration with several academic institutes.

The model is inspired by a real Cypriot coastal city but has been appropriately reworked for security reasons and made available in EPANET format as an open-benchmark network. It is now an integral part of the KIOS Virtual City Testbed, a platform that allows the coordinated simulation of critical infrastructures (water, electricity, transport and telecommunications) in realistic operational and crisis scenarios.

The availability of such a model stems from the need to have shared, transparent and replicable case studies, which allow a fair comparison between different methods of leakage diagnosis. In the scientific field, in fact, the performance of detection methods is often difficult to evaluate objectively: the real datasets of water companies are generally not accessible, partial or affected by undocumented noise and uncertainties, while synthetic models are often too simplified and lacking in engineering meaning.

L-Town fills this gap, providing an "urban-scale" system in which each parameter is defined with realistic but controlled criteria, and in which operational uncertainties (e.g. demand variability, errors on the piping parameters, unknown valve status) are explicitly modeled.

In the context of the *BattLeDIM* competition, the model was used to test very different approaches, from methods based on physical models and hydraulic residues to machine learning and signal processing algorithms, all evaluated through a single economic scoring function that rewards the ability to identify leaks early and accurately. while minimizing the operational costs of research and repair.

Clarification on the adoption of the L-Town network in this work

L-Town has been chosen in this thesis as the main reference case, first of all, because it is a benchmark widely recognized and documented in recent literature, as explained above.

However, it should be noted that in this work the *L-Town* network is not used for the purpose of direct comparison with the results of the *BattLeDIM* competition or other studies in literature. In particular, the goal of this research is not the development of new leak detection or localization techniques, but the proposal of a chain of integrated processes oriented to the mitigation of the risk from HDL (Hydrogeological Disruption due to Leakage).

Within this supply chain, the leak localization phase represents only one of the constituent modules and is implemented using methodologies already consolidated in the literature. For this reason, the *L-Town* network is used not to innovate the location itself, but to illustrate in a more understandable way how the overall process proposed in this thesis works.

In this perspective, the leak datasets released in the context of the *BattLeDIM*, since our goal is different: while in the competition the leaks are simulated only in a limited number of sections of the network, in our case we intend to train and optimize the positioning of a limited number of sensors to be able to locate leaks potentially in any pipe of the system and in particular in the areas with increased risk from HDL.

For this reason, in the following sections we will describe in detail how a dedicated dataset was generated, built using only the topological structure and the hydraulic model of the network (derived from the *L-Town.inp* file) and leaving out the rest of the data that is provided for the purposes of the Battle.

Consequently, the results in terms of localization accuracy reported below do not constitute the main focus of the research but represent exclusively a functional verification of the integrated process.

It is important to emphasize that adopting a different leak localization methodology leads to different accuracy values; however, this does not represent a limitation of the proposed framework. From this perspective, other methods, either among those proposed in the *BattLeDIM* or available in the literature, can also be employed, since one of the main strengths of the developed framework lies in its modular design, conceived as an integrated sequence of processes that can be combined according to a coherent and flexible logic

Therefore, the innovation of the present work does not lie in the localization technique itself, but in the systemic and modular approach that integrates localization with other processes (risk zoning, optimization, etc.) aimed at mitigating HDL risk.

In this perspective, the *L-Town* network represents an ideal case study for the demonstration of the proposed method, thanks to its wide diffusion, public availability and the well-established structure in the scientific community, elements that facilitate its application, replicability and future extension.

Topology, scale, and hydraulic components

The *L-Town* model is implemented through EPANET 2.2 [48] with a level of detail comparable to that of a real medium-sized urban network (Figure 30). It consists of 782 junctions and 905 pipelines, with an average length of 50 m each, for a total extension of 42.6 km. All ducts are made of steel with Hazen–Williams roughness coefficients between 120 and 140. The entire network serves about 10,000 users between civil and industrial and has a loop ratio of 25%, i.e. a quarter of the pipelines would have to be removed to eliminate all the meshes (Vrachimis et al. 2019) [108]. The elevations of the nodes vary between 1.5 m and 75 m a.s.l.

The network is fed by two main tanks, which guarantee a pressure of between 20 and 30 m under normal operating conditions. A pressure reducing valve (PRV) is installed in the lower part of the city (Area B), intended to limit bottom leaks, while other PRVs are located downstream of the tanks. The upper part of the city (Area C) is served by a pump and a cylindrical tank of 16 m in diameter, which fills up at night and supplies water during the day. This configuration faithfully reproduces the typical behavior of many coastal cities characterized by pronounced elevation differences (see Figure 30).

The choice to segment the pipelines into 50 m sections was deliberate: on the one hand it reflects the average distance between real users, on the other hand it simplifies the management of leaks in the benchmark, allowing them to be uniquely identified at the level of the individual pipeline. In addition,

this discretization allows users of the model to perform any complexity reductions on their own, while still being able to compare the results with the full version.

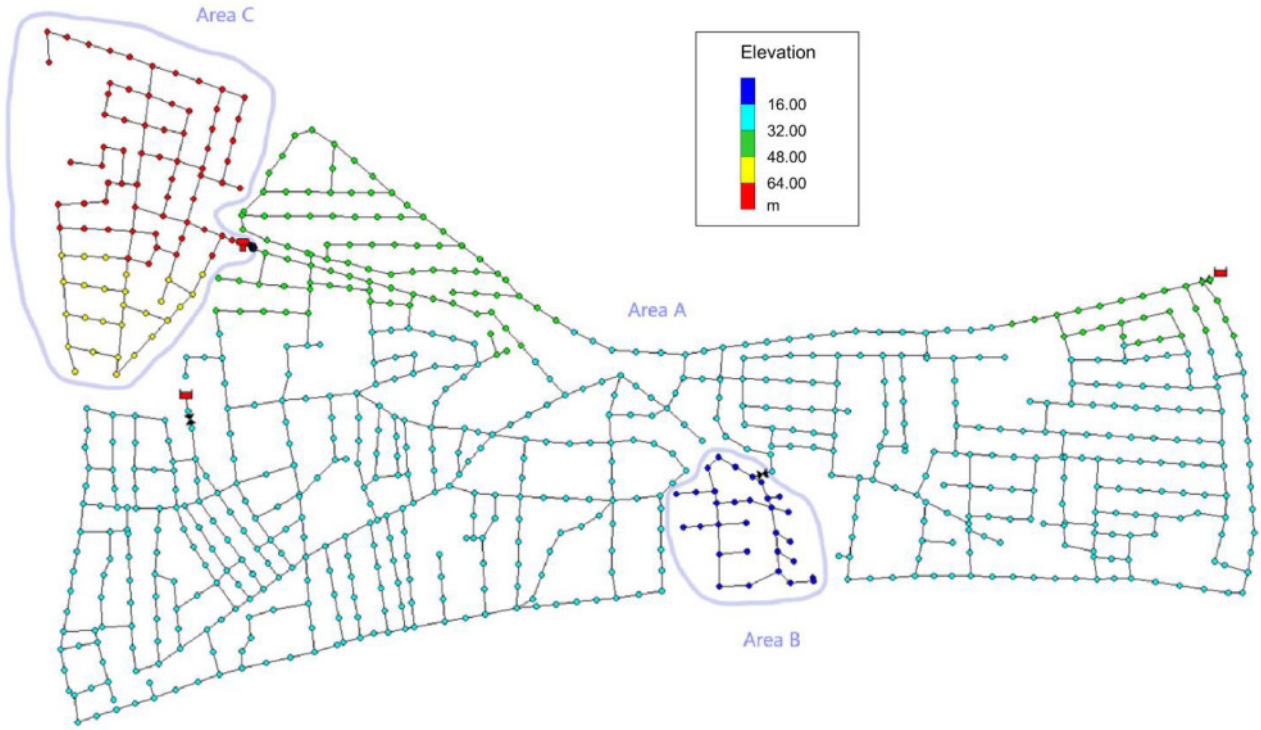

Figure 30. Reproduced from Vrachimis et al. (2022) [107]. Altimetric subdivision of the L-Town water network. Nodes are colored according to piezometric elevation, with increasing values from blue (16 m) to red (64 m). The network is divided into three macro-areas: Area B, corresponding to the lowest elevation and characterized by the lowest elevations; Area A, of intermediate elevation and morphologically central; and Area C, representing the highest portion. This altimetric distinction highlights the topographic variability of the system, useful for hydraulic analysis and monitoring planning.

Measurement system

The *L-Town* network is equipped with a remote control system (SCADA) consisting of a limited but representative number of sensors, capable of reproducing the typical configuration of a medium-sized real network:

- 1 level sensor in the tank;
- 3 flow sensors (located at the pump and at the DMA entrances);
- 33 pressure sensors, distributed according to a hydraulic sensitivity criterion (see Figure 31);
- 82 Automatic Meter Reading (AMR) devices in Area C, which provide aggregate consumption data.

The sensors transmit readings every 5 minutes, with no delays or packet loss. Pressures are averaged over 5-minute intervals, to mitigate the effects of transients, and data is rounded to two decimal places, in line with real-world SCADA systems. This approach allows the evaluation of methodologies to be focused on analytical capabilities, excluding effects related to noise or data loss.

The available datasets include two years of simulations (2018 and 2019): the first represents the "historical" year, containing some known and subsequently repaired leaks, while the second constitutes the "evaluation" dataset, used to compare the performance of the different localization methods.

In this work, however, the original SCADA data were not used, but only the positions of the 33 pressure sensors proposed in the *BattLeDIM*, used as potential positions for defining the set of sensor nodes to be optimized.

The number of sensors actually adopted in this study is considerably less than 33, in order to approach a more realistic and representative value of the configurations commonly found in real operational applications.

The configuration of the sensors proposed in the *BattLeDIM*, obtained by the sensitivity matrix method, is shown in the following figure.

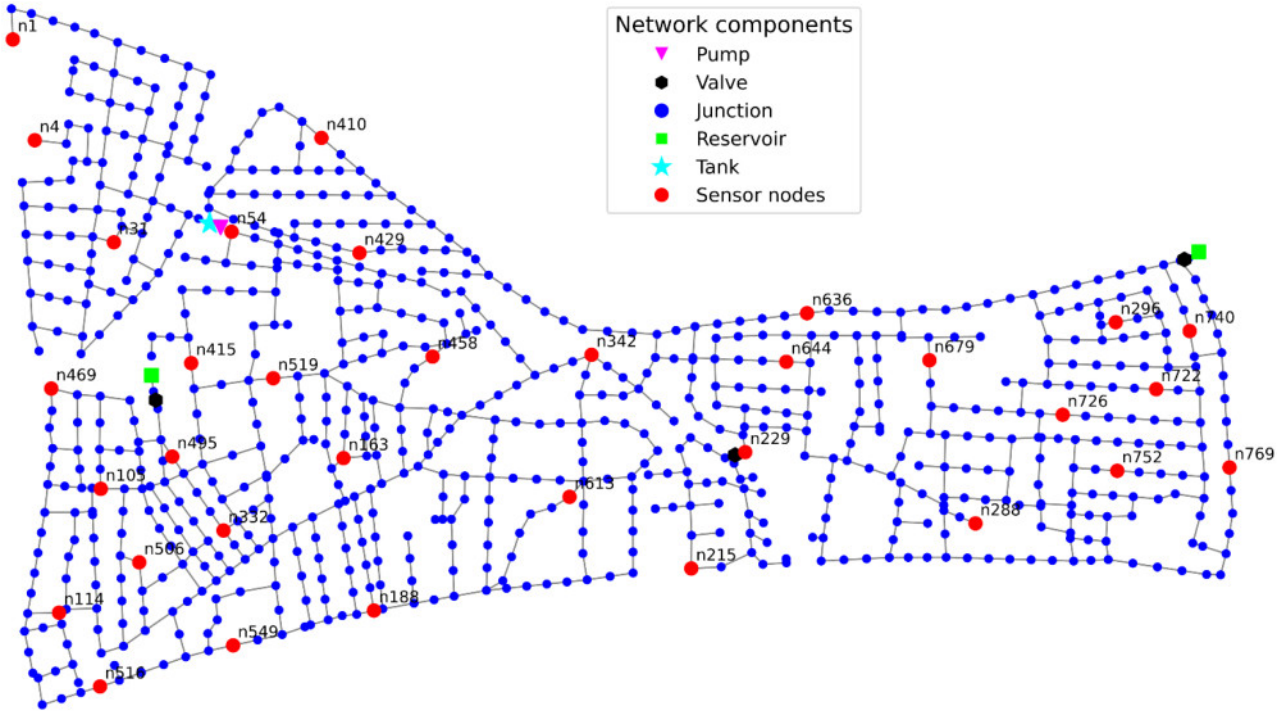

Figure 31. Schematic of the L-Town water network. The map shows the system's main components. The 33 pressure sensors were placed using the sensitivity matrix method.

Model uncertainties and differences between real and nominal network

A crucial aspect of the *BattLeDIM* benchmark is the distinction between the nominal model (provided to participants) and the real model (used to generate the reference data). The nominal model faithfully reproduces the topology and general structure of the network, but includes some controlled uncertainties, including:

- random variations (±10%) on key hydraulic parameters, such as roughness, pipe diameter and length;
- disruptions to basic demand and consumption patterns;
- absence of specific industrial patterns;
- two pipelines (p37 and p251) closed in the real model but open in the nominal model.

These differences force localization algorithms to manage the discrepancy between simulation and reality, realistically reproducing the challenges faced by water utilities, where mathematical models never coincide perfectly with actual operating conditions.

In this work, however, only the nominal model was used, for the reasons previously explained: the objective is not to evaluate the performance of a specific localization method, but rather to propose and test an integrated process chain aimed at mitigating HDL risk. To this end, data generated by the real model were not used, but only those from the nominal model for demonstration purposes, simulating different leak scenarios using the EPyT and WNTR libraries. These simulations allowed the sensors to be trained to recognize and locate leaks across the entire network, and not just at the points covered by the benchmark.

Looking ahead, further studies could integrate the proposed framework with the *BattLeDIM* approach, leveraging its real datasets to achieve an even more comprehensive validation.

Scientific and operational implications

Therefore, compared to the previous case, the *L-Town* benchmark stands out for its transparency and controlled complexity.

In the framework of this thesis, *L-Town* is used as a reference platform for three main aspects:

1. Carry out the zoning of the risk from HDL, in order to identify the different levels of risk;
2. Generate a leak scenario dataset, where each scenario corresponds to a localized leak on a single pipeline, covering all pipelines in the network;
3. Optimize sensor placement, taking into account zoning and pursuing HDL risk mitigation, by maximizing location accuracy in areas of higher risk.

Summary of the characteristics of the L-Town functional to this study

- Components: 782 nodes; 905 pipes of 50 m; 42.6 km in total; ~10.000 users;
- Pipe diameters: ranging from 63 to 225 mm;
- Power and controls: 2 tanks; PRV (incl. Area B); 1 pump + tank (Ø ≈16 m) for Area C;
- Operating pressures: 20–30 m;
- Demands:
    - built from GIS data + Fourier series;
    - PDD;
    - minimum pressure threshold, $P_0 = 7\text{m}$; full demand pressure, $P_f = 25\text{m}$;
    - pressure exponent, $\delta = 0{,}5$; demand peaking factor (DPF) = 1,5–2,0;
- Telemetry: 33 pressure sensors;
- Uncertainties: ±10% on parameters and demands.

3.2.2. L-Town HDL Risk Zoning

In the case study of the *L-Town* network, the zoning of the risk of hydrogeological disruption caused by leaks (HDL) was carried out using a more structured approach than that adopted for *Real Network 1*. In this case, in fact, not only the exposure component (E) was considered, but also the hazard component (H), while vulnerability (V) was kept constant and homogeneous throughout the analysis domain, as a simplifying assumption, since sufficient data are not available for its accurate assessment.

The overall risk assessment was therefore carried out through a combination of the two components assessed, which it was decided to weight by weights to reflect the different reliability and completeness of the available data. In particular, a greater weight was attributed to exposure, as it was evaluated in more detail and supported by more solid territorial and demographic information, while the hazard component was only partially estimated.

The model adopted is expressed in the general form:

$$R(x) = \alpha\, E(x) \circ \beta\, H(x) \tag{23}$$

where the symbol "∘" indicates a generic operation of combination between the two components and the coefficients α and β represent the relative weights assigned respectively to exposure and hazard. This formulation allows the model to remain general and flexible, without constraining it to a specific operational form, while preserving its consistency with the nature and level of detail of the available data.

The assessment of territorial exposure was conducted through the integration of demographic and territorial data from official sources and satellite observations, similar to the previous case.

Demographics

Population data were obtained from the *CityPopulation.de* portal [109], which provides up-to-date information on population density and population distribution for each administrative section of the district under study, in whose territory the *L-Town* network is located.

The portal is based, in turn, on official data provided by the Statistical Service of Cyprus (CyStat) [110], the national body responsible for the collection and dissemination of demographic and territorial statistics for Cyprus. This information made it possible to accurately estimate the population density in the urban areas crossed by the network, identifying the areas with the greatest population concentration.

It should be noted that the exact geographical location of the network cannot be made public, as it is not made explicit in official sources. However, through a territorial and cartographic analysis it was possible to identify the area in which the network develops. The demographic and territorial data used therefore actually refer to the context of *L-Town*, but, in line with the literature and with the confidentiality practices of public datasets, the precise position is not disclosed.

Land use and land cover

The Urban Atlas Land Cover/Land Use 2018 dataset (vector), published by the European Environment Agency (EEA) as part of the Copernicus Land Monitoring Service (CLMS) [111], was used for the characterisation of land use.

The dataset, an integral part of the European Union's Copernicus programme, represents an official, harmonised and high-resolution source of information on land cover and land use in Europe's main urban areas.

The Urban Atlas 2018 provides reliable, comparable and consistent data for 788 Functional Urban Areas (FUAs) with more than 50,000 inhabitants, referring to the year 2018, and covers the EEA-38 countries (EU Member States, EFTA countries, Western Balkans and Turkey) as well as the United Kingdom.

The dataset provides vector data with a spatial resolution of 10 meters, ensuring high accuracy in the representation of spatial planning and the degree of anthropization. Land use/cover classes are organized in a multi-level thematic hierarchy and include 17 urban classes, with a minimum mapping unit of 0.25 ha, and 10 rural classes, with a minimum mapping unit of 1 ha.

Access to the data is completely free of charge, in accordance with the Commission Delegated Regulation (EU) No. 1159/2013, which establishes the principle of open and free access to Copernicus programme products. The only obligation for users is to cite the source and declare any changes made to the original data.

Satellite observations

As in the case of *Real Network 1*, the information derived from Urban Atlas and *CityPopulation.de* has been verified and updated through photointerpretative analysis on high-resolution satellite images, obtained from Google Earth Pro and Google Maps. This has made it possible to correct any discrepancies due to recent transformations of the territory, such as new urbanizations, infrastructures or changes in intended use.

Classification criterion

In the present study, the Urban Atlas 2018 dataset was used to derive the territorial component of the exposure (E), starting from the land use categories and associating them with the use classes provided for by the Technical Standards for Construction (NTC 2018).

In other words, the NTC 2018 has been taken as a reference criterion for the functional classification of the territory, allowing the categories of the Urban Atlas to be interpreted from a structural and infrastructural point of view, depending on the importance of the settled activities and the degree of anthropization.

On the basis of this criterion, the works are divided into four classes of use (I–IV), defined in relation to strategic importance, the level of crowding and the consequences of any interruption or collapse. As in the previous case, classes III and IV have been merged into a single category, thus obtaining three levels of exposure (E1-E3) consistent with the territorial structure and with the functional role of the different areas.

In order to integrate the spatial information of the Urban Atlas 2018 with the criterion defined by the NTC 2018, a direct correspondence was made between the Urban Atlas encodings and the use classes of the technical standards.

Each type of land use/cover of the Urban Atlas, identified by a unique code and associated with specific territorial portions, was analyzed according to its function and nature.

On the basis of these characteristics, each coding has been traced back to the corresponding NTC 2018 use class, ensuring a consistent classification between the land use and the structural and functional importance of the areas represented.

Table 7 shows the defined correspondence, used to attribute to each urban section an equivalent level of exposure (E1–E3), proportionate to the exhibition and functional value.

Table 7. Correspondence between the Urban Atlas 2018 land use/cover classes and the NTC 2018 use classes, used for the definition of exposure levels (E1–E3) in the study domain of the L-Town network.

NTC 2018 Use Class	Level of exposure value	Code and description of the types of land cover/use of the UA 2018
I	Low (E1)	11240 - Discontinuous Very Low Density Urban Fabric (<10%) 11300 - Isolated Structures 13400 - Land without current use 14100 - Green urban areas 21000 - Arable land (annual crops) 22000 - Permanent crops (vineyards, fruit trees, olive groves) 23000 – Pastures 24000 - Complex and mixed cultivation patterns 25000 - Orchards at the fringe of urban classes 31000 – Forests 32000 - Herbaceous vegetation associations 33000 - Open spaces with little/no vegetation 40000 – Wetland 50000 - Water bodies 13100 - Mineral extraction and dump sites
II	Medium (E2)	11220 - Discontinuous Medium Density Urban Fabric (30–50%) 11230 - Discontinuous Low Density Urban Fabric (10–30%) 12220 - Other roads and associated land
III, IV	High (E3)	11100 - Continuous Urban Fabric (>80%) 11210 - Discontinuous Dense Urban Fabric (50–80%) 12100 - Industrial, commercial, public, military and private units 12210 - Fast transit roads and associated land 12230 - Railways and associated land 12300 - Port areas 12400 – Airports 13300 - Construction sites 14200 - Sports and leisure facilities

Exposure components

The exposure component has been constructed from three main sub-elements, combined in a weighted way to obtain an overall synthetic value:

- Qualitative exposed value of built assets;
- Population density;
- Importance of road network.

The qualitative exposed value of built assets represents the density and importance of the building stock, accounting for the economic and social impacts associated with structural damage or collapse.

For the case in question, exposure value zoning was performed across the entire urban area affected by the network, considered as a continuous domain (Figure 32). Subsequently, this zoning was intersected with the network buffer, in order to transfer the exposure values only to the areas located in proximity to the pipelines (Figure 33).

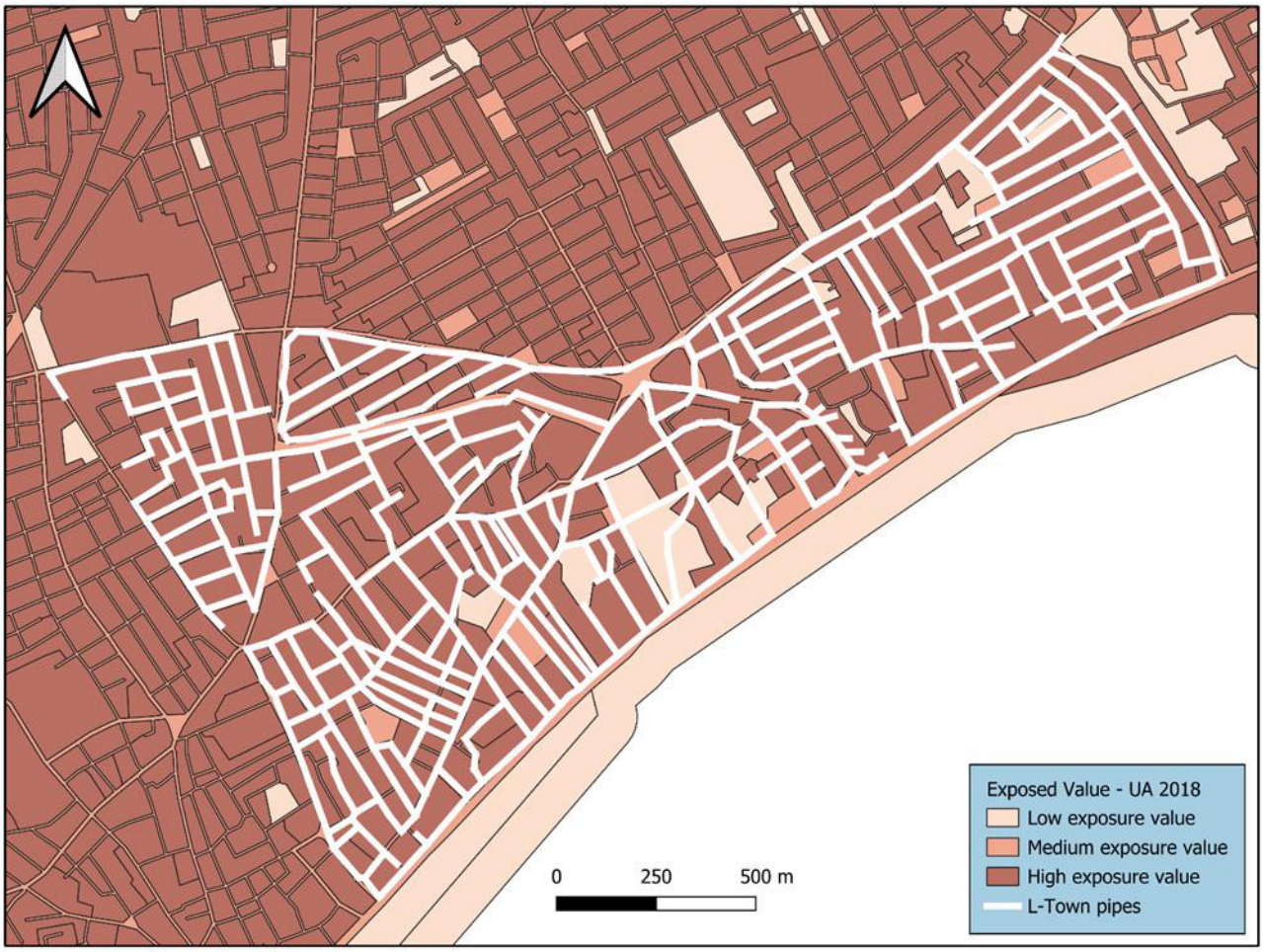

Figure 32. Zoning of the exposure value of built assets over the entire urban area affected by the L-Town network.

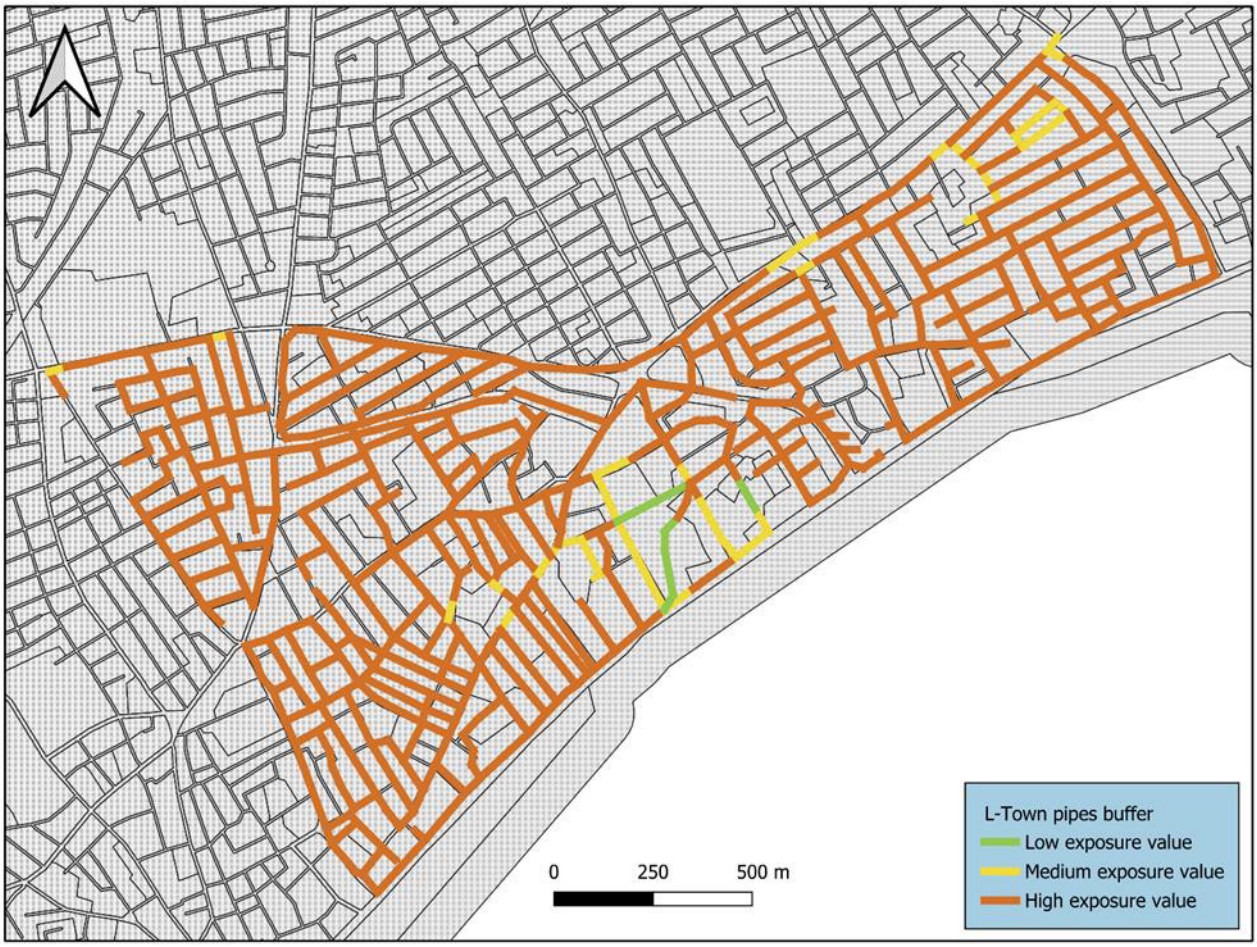

Figure 33. Zoning of the exposure value of built assets for the L-Town network buffer zone

The population density represents the concentration of the population residing in the areas surrounding the network.

This component is fundamental, as it constitutes the very essence of the exhibition value: in the absence of human presence, in fact, the risk associated with an event would be limited. As a result, the most densely populated areas are also potentially the most critical in terms of exposure.

In this case too, a process similar to that adopted for the zoning of the exposure value of real estate was followed: initially, a continuous analysis was conducted on the entire urban area covered by the network, considering the different census sections and the relative distribution of the population (Figure 34).

Subsequently, this information was transferred in synthetic form to the network buffer, in order to geolocate the population density values in correspondence with the pipelines (Figure 35).

In other words, the population density (inhabitants/km²), originally referred to the territorial sections, has been projected and aggregated along the route of the network through a process of spatial intersection, so as to coherently represent the actual anthropogenic exposure in the areas close to the pipes.

This operation makes it possible to trace all exposure information directly to the elements of the water network, ensuring full geographical consistency between the different components (real estate, population density and road network).

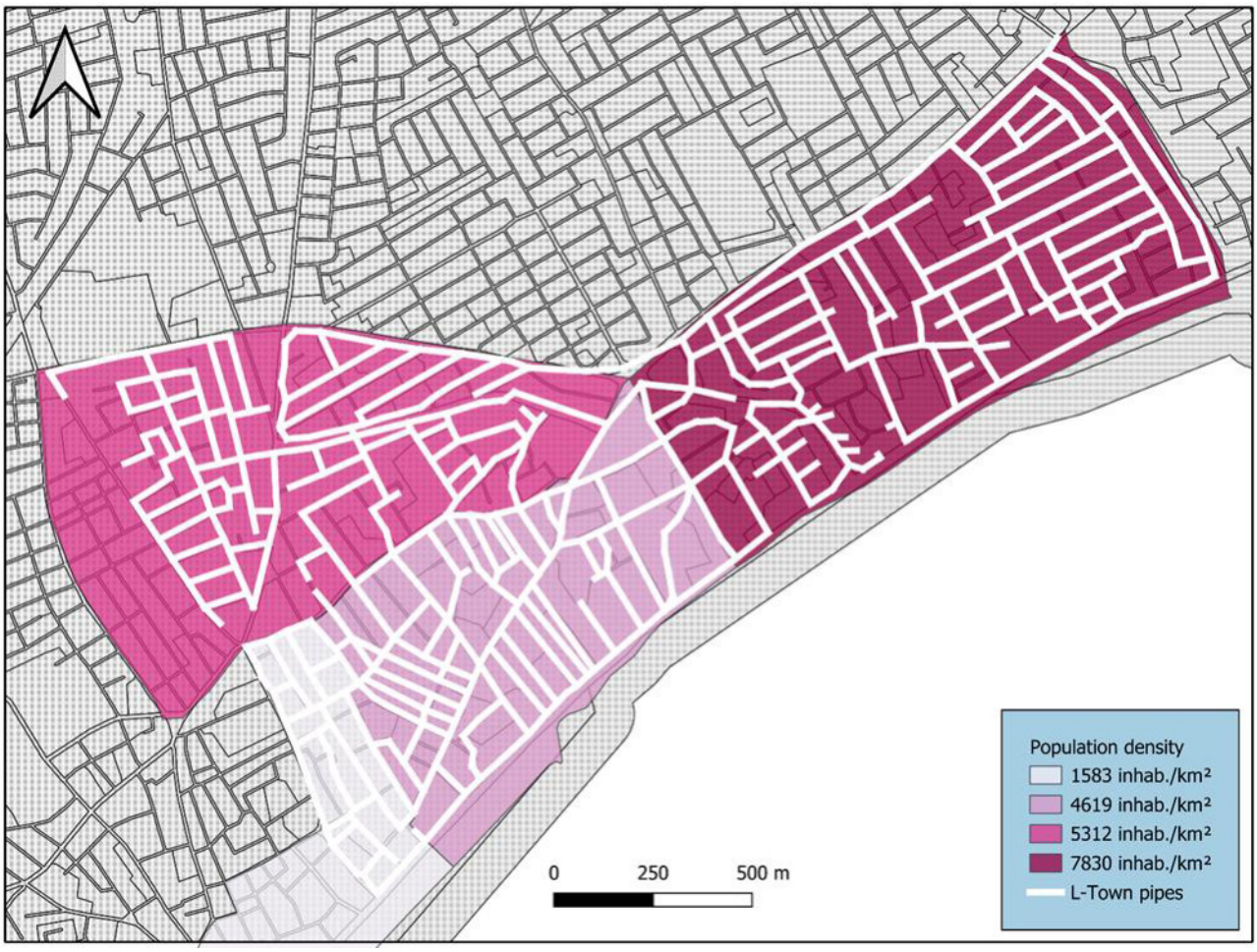

Figure 34. Zoning of population density over the urban area affected by the L-Town network.

This spatial harmonization is essential, since in the next phase the various components will have to be combined to obtain a synthetic indicator of exposure: for this reason, each data must refer to the same geographical domain, i.e. to the pipes of the network.

The importance of the road network assesses the degree of functional relevance of the road network in the vicinity of the pipelines.

Main roads and thoroughfares are more impacted, as their closure would have more serious consequences for urban functioning; the impact on secondary roads is less severe, and even less so on dead ends, where the consequences of potential road damage are more limited.

The water distribution network generally develops along the route of the roads, as this arrangement facilitates the installation, maintenance, inspection and monitoring operations. For this reason, the assessment of the importance of the road network does not require a preliminary territorial analysis of the entire urban domain but can be carried out directly by following the route of the pipes, to which the related road infrastructures are already associated.

Once the correspondence between pipelines and roads has been identified, the latter are classified according to their functional importance (Figure 36).

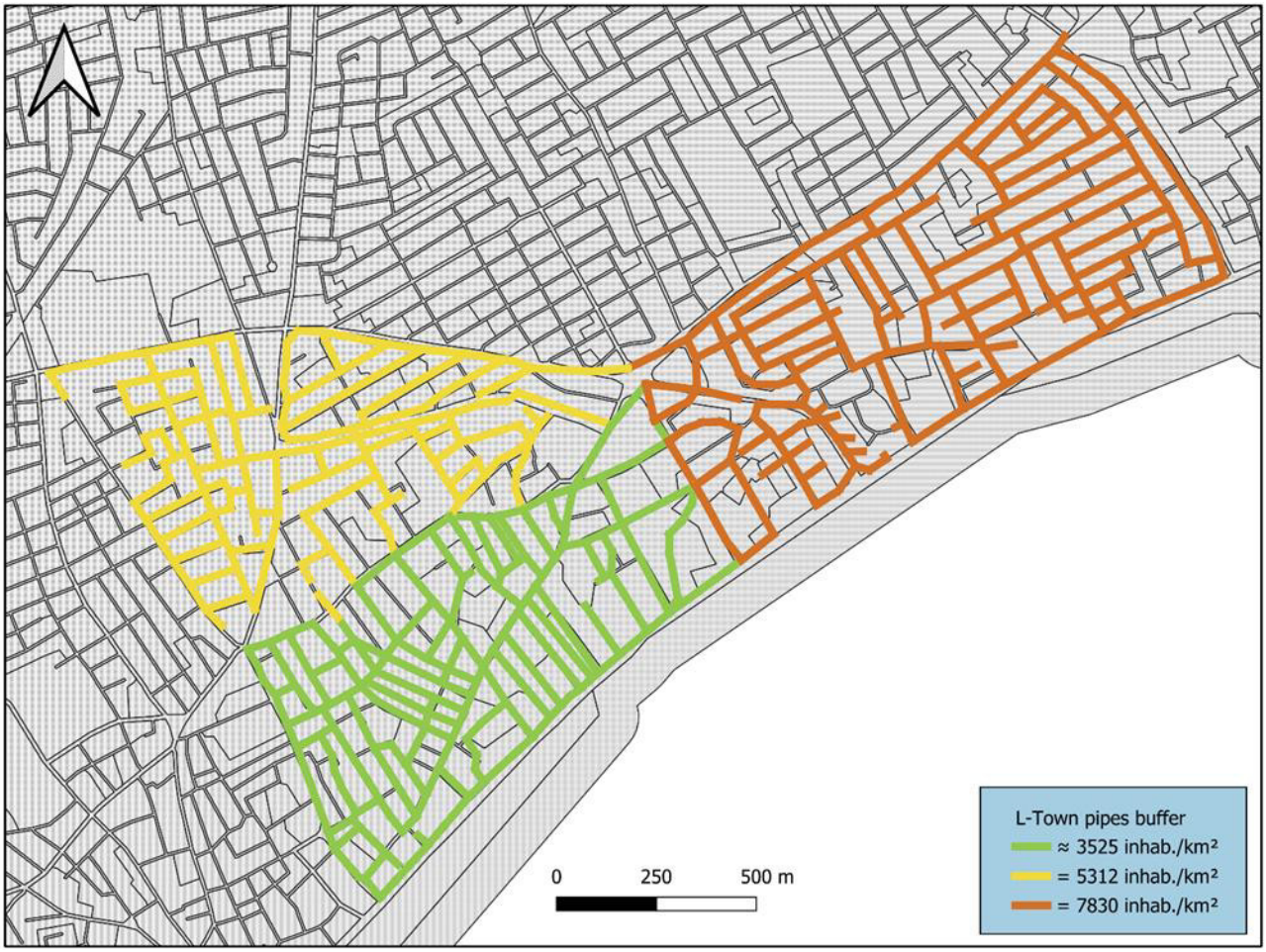

Figure 35. Zoning of population density for the L-Town network buffer zone.

The main arteries are considered the most important, as they support greater flows of vehicular and pedestrian traffic, frequently host buildings and services of public importance and play a strategic role for the continuity of the urban fabric and the local economy. Secondary roads are of intermediate importance, as they are characterised by moderate traffic.

Finally, local streets or dead ends are frequented almost exclusively by residents or regular users, and therefore the impact of any damage due to leaks is limited and confined to the surrounding area.

HDL Risk Exposure Component

The three components (qualitative value of real estate, population density and importance of road network) were subsequently combined (Equation 24) using weights calibrated according to their relative incidence, in order to represent in a balanced way the contribution of each factor to the overall exposure.

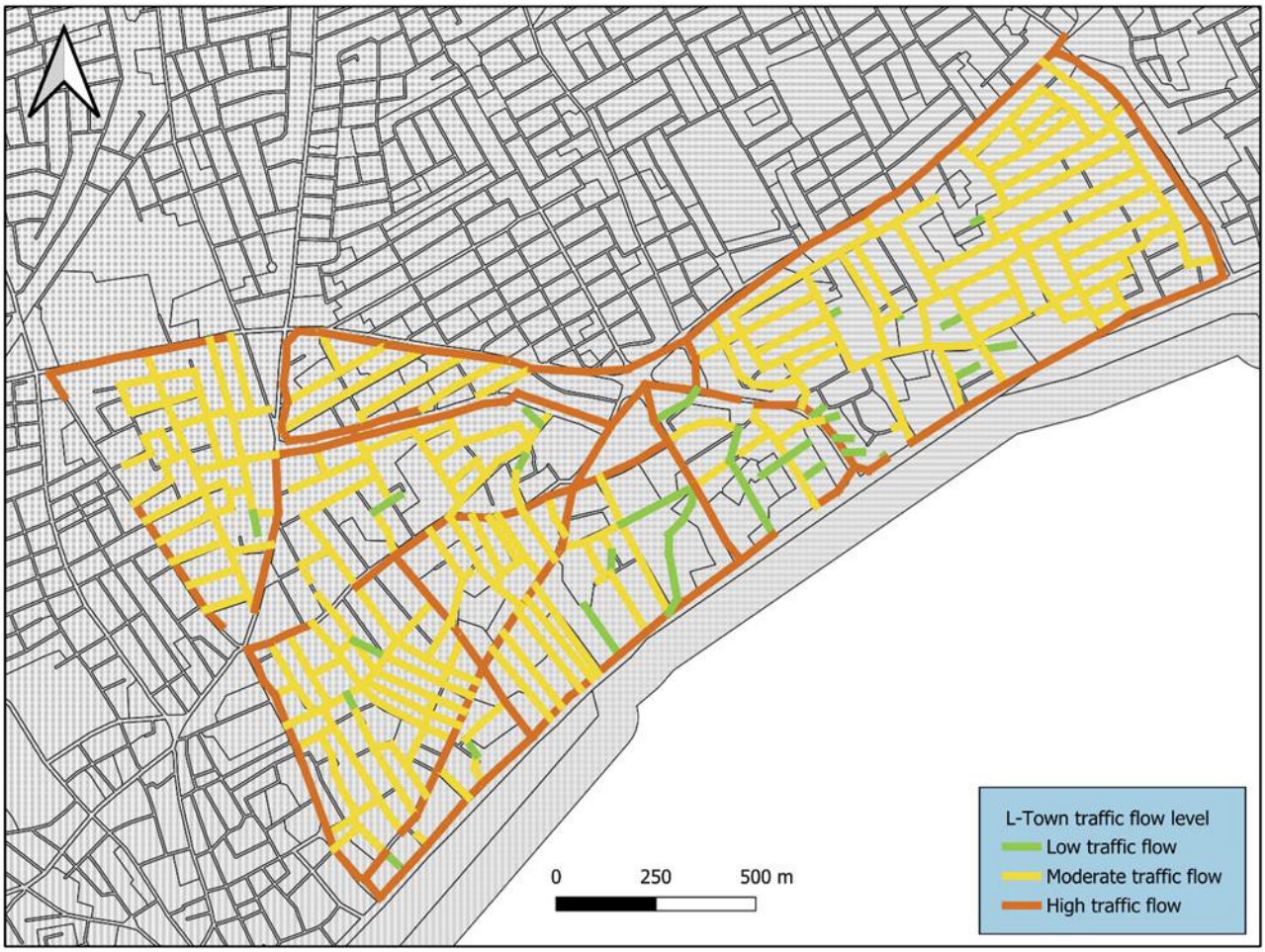

Figure 36. Zoning of road network importance near the L-Town pipelines, obtained by classifying roads according to their functional role: main, secondary, and local (for example, dead-end streets).

$$E(x) = \frac{w_{ba}\, E_{ba}(x) + w_{pd}\, E_{pd}(x) + w_{rs}\, E_{rs}(x)}{w_{ba} + w_{pd} + w_{rs}} \tag{24}$$

where:

- $E_{ba}(x)$ the exposed qualitative value of built assets;
- $E_{pd}(x)$ the exposure in terms of population density;
- $E_{rs}(x)$ the importance of the road system.
- w_{ba}, w_{pd}, w_{rs} are the weights associated with the respective components, determined in such a way as to reflect their relative impact on the overall territorial exposure.

The result of the combination is an integrated exposure map (E), which synthesizes demographic, structural and socio-infrastructural information, providing a coherent representation of the areas with different exposure values with respect to HDL risk (Figure 37).

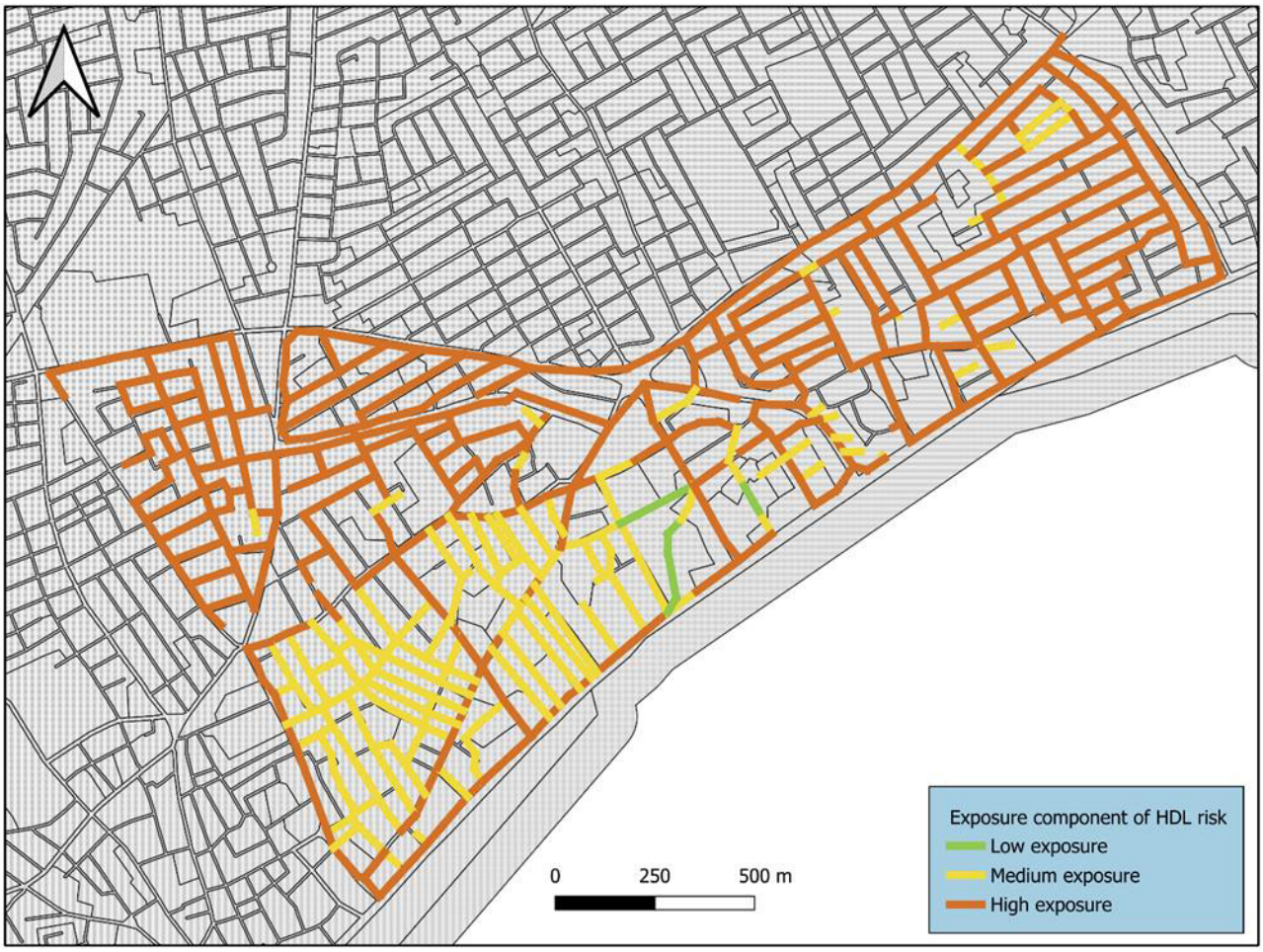

Figure 37. Overall HDL risk exposure map (E) for the L-Town network.

Hazard Components

The hazard component (H) was assessed taking into account the information available for the *L-Town* network, with the aim of representing, as far as possible, the pipeline's propensity for failure.

In general, a comprehensive hazard assessment would require the availability of extensive information, not only operational, but also intrinsic and environmental. In addition to hydraulic parameters, data regarding pipeline material, installation depth, soil conditions, infrastructure age, and other factors that influence the likelihood of leaks would be required.

In the present case, however, only some information is available, in particular relating to:

1. *Pipe diameter*: For certain materials, such as steel, smaller diameter pipes are more prone to breakage and therefore have a higher level of hazard.
2. *Operating pressure*: according to the literature of the sector, this is one of the main risk factors, since high pressure values increase the stress of the pipelines and consequently the probability of failures and leaks.

Again, the two contributions have been combined in a weighted manner, giving more weight to the operating pressure, as it is recognized that operating conditions more significantly influence the probability of failure.

However, since the assessment does not include other structural and environmental factors, the hazard component was considered with a lower overall weight than the exposure component in the definition of the final risk.

Failure rate as a function of pipe diameter

As reported in Barton et al. (2019) [40], there are numerous studies in the literature that have highlighted a strong relationship between the rupture rate of pipelines and their diameter.

In general, pipelines with a diameter of less than 200 mm have the highest failure rates, a phenomenon attributable to several factors including lower resistance to ground movements and corrosion (reduced thicknesses), poor reliability of the joints, and greater exposure to excavation activities or urban vibrations due to the more superficial laying (Gould et al., 2013; Bruaset and Sægrov, 2018). [113, 114]

In the case of the *L-Town* network, the pipes are made of steel, therefore, for the assessment of the hazard as a function of the diameter, the curve corresponding to the steel & ductile iron shown in Figure 38 was taken as a reference.

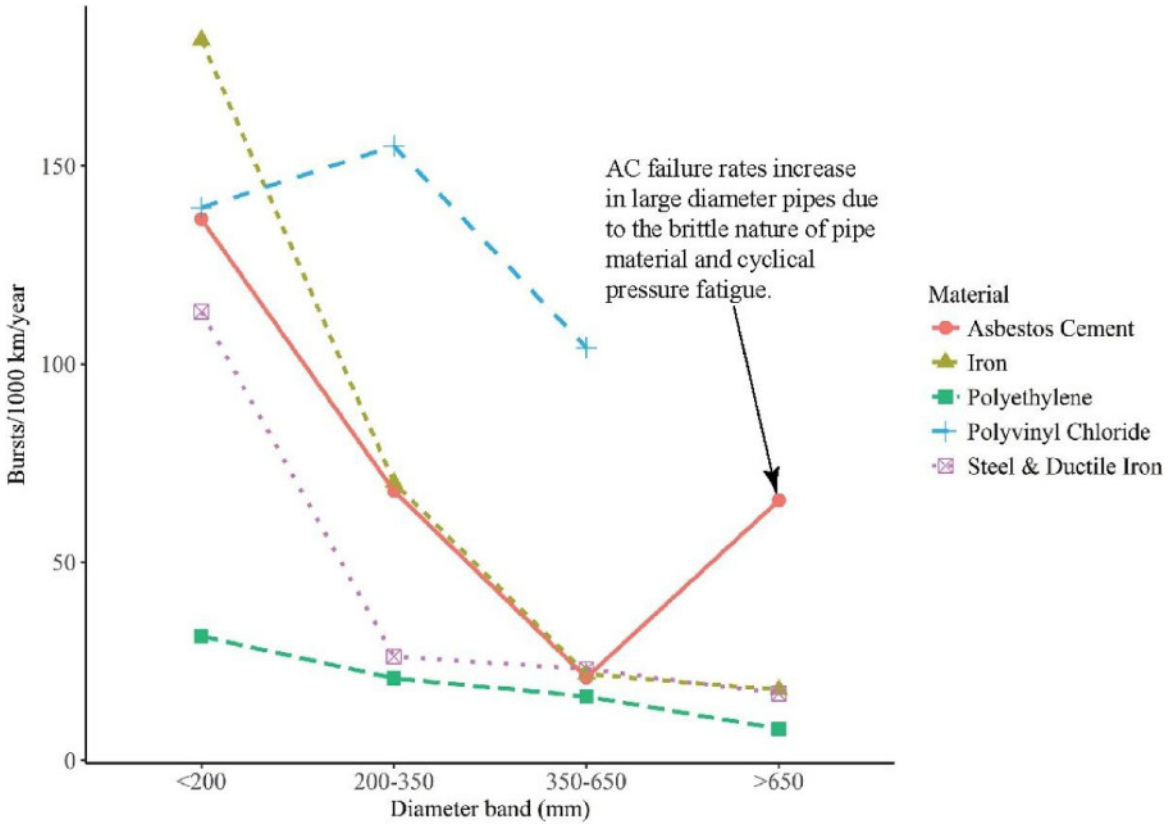

Figure 38. Reproduced from Barton et al. (2019) [40]. Relationship between pipe failure rates and diameter size for different material types.

Based on this relationship, the diameters of the network have been divided into three dimensional ranges, associating increasing levels of danger with smaller diameters, which are statistically more prone to breakage.

Figure 39 shows the hazard map deriving from the diameter, in which the pipelines are classified according to these ranges, reflecting the different propensity to failure depending on their section.

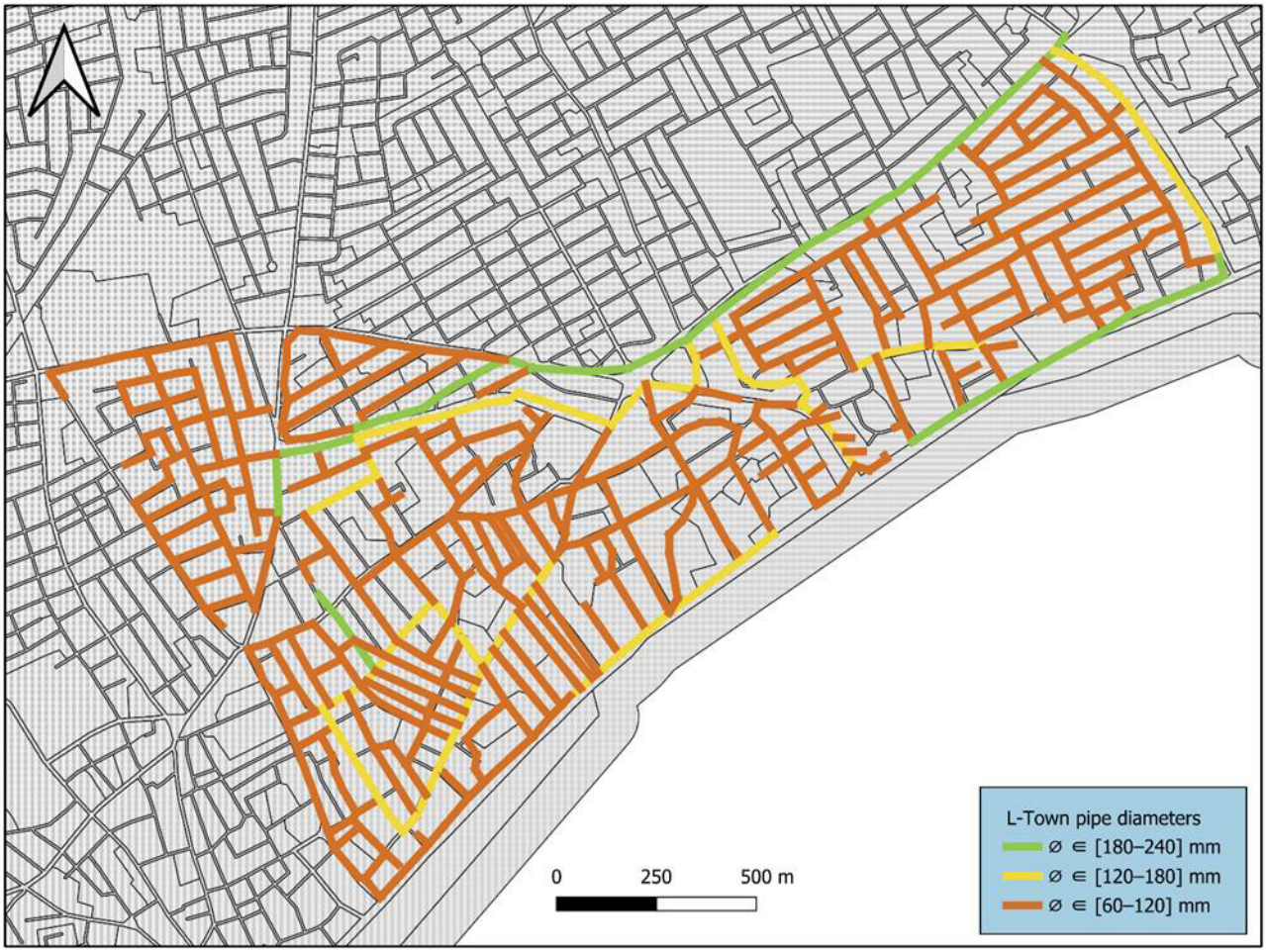

Figure 39. Hazard map derived from pipe diameter for the L-Town network

Zoning of the network according to operating pressures

Operating pressure is one of the main factors influencing the frequency of pipe ruptures in water distribution networks. It is widely recognized in the literature that high pressure values or a significant variability of the same over time can lead to an increase in mechanical stress on the pipes, favoring fatigue phenomena and the subsequent onset of failures. For this reason, the zoning of the network according to operating pressures is a useful tool both for assessing the probability of structural damage to the sections and for setting up pressure management and preventive maintenance strategies.

Martínez-Codina et al. (2015) [115] showed that indicators derived from pressure time series are among the most predictive of the probability of breakup. The comparison between the cumulative rupture-conditional distribution and random samples made it possible to highlight that the pressure range is the most representative indicator, confirming that not only the maximum values but also the oscillations contribute decisively to failures.

Instead, Moslehi and Jalili-Ghazizadeh (2020) [116] proposed a field-based methodology to statistically describe the relationship between pressure and rupture rate, developing Break Rate

Functions (BRFs) for different conduction materials. The approach is based on the use of a Maximum Pressure Indicator (MPI), calculated in the midpoints of the area, and allows to identify, through Bayes' theorem, specific pressure thresholds beyond which a sudden increase in the frequency of ruptures is observed.

More recently, Konstantinou et al. (2023) [117] introduced the concept of Cumulative Pressure-Induced Stress (CPIS), an index that integrates the average pressure, amplitude, and frequency of pressure cycles over time, in order to estimate the cumulative stress and fatigue risk of pipelines. The inclusion of CPIS in machine learning (Random Forest) models has improved the predictive capacity of the models, confirming that pressure-related parameters play a decisive role in predicting failures.

In the case of the network analyzed, zoning was carried out considering the pressure ranges within which the pipelines operate. Although the oscillations may have smaller amplitudes than those represented in the zoning, it was decided to distinguish two main pressure ranges: one between 25 m and 50 m of water column, which mainly affects the part of the network facing inland, and one between 50 m and 60 m, corresponding to the area closest to the coast (see Figure 40).

These values are, however, consistent with the regulatory and operational references commonly adopted in urban water systems.

In Italy, the technical regulations of the Integrated Water Service operators (e.g.: CAP Group [118]; SMAT [119]) establish that:

- the minimum guaranteed pressure varies between 5 m (relative to the roof slab of the highest habitable floor of the building served [119]) and 20 m [118] of water column, depending on the height of the user and local conditions;
- the maximum operating pressure must not exceed a value between 70 (referred to the delivery point, in relation to the road surface [119]) and 100 m of water column [118] (≈ 10 bar).
- pressures between 30 m and 60 m (≈ 3 – 6 bar) are generally considered optimal to ensure efficient service and reduce the risk of breakage.

Therefore, the choice to divide the network into two macro-pressure classes is dictated more to differentiate different stress conditions, being normal operating pressures.

- il first interval (25–50 m) represents an optimal operating range, consistent with the minimum limits and with most Italian and European regulations;
- the second range (50–60 m) identifies a higher pressure band, potentially more vulnerable to the risk of ruptures, in line with experimental evidence. However, it is a more depressed area and therefore in the long term it will tend to always be subject to higher pressures than other parts of the network.

In conclusion, the zoning of the network on the basis of operating pressures makes it possible to identify operating areas with different degrees of hydraulic stress, providing objective support for the planning of regulation and maintenance interventions. It therefore represents a key step in the process of mitigating the risk from HDL.

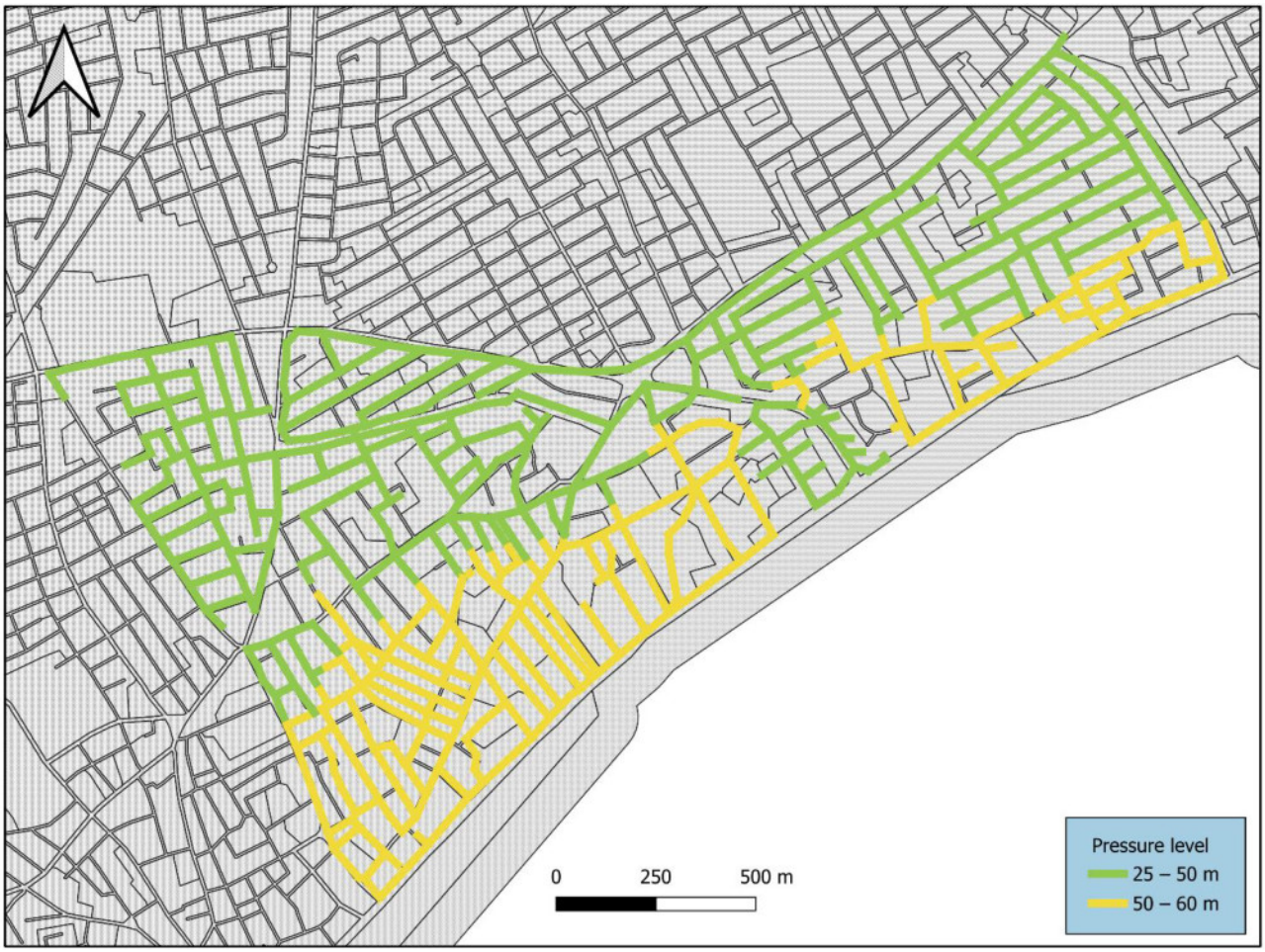

Figure 40. Distribution of pressure levels in the L-Town network: pipelines with pressures between 25 and 50 m of water column (in green) mainly affect the western and inland portion of the system, while those between 50 and 60 m (in yellow) are more concentrated in areas close to the coast.

HDL Risk Hazard Component

Once the two hazard components have been defined, (the structural one, linked to the diameter of the pipes, $H_D(x)$, and the operating one, associated with the operating pressure levels, $H_P(x)$), the overall Hazard index (H) was calculated by means of a weighted combination of the two.

In order to give greater importance to operating conditions, weighting coefficients w_D and w_P have been introduced, with $w_P > w_D$, according to the following:

$$H(x) = \frac{w_D\, H_D(x) + w_P\, H_P(x)}{w_D + w_P} \tag{25}$$

The resulting hazard zoning map (Figure 41) highlights the spatial distribution of the propensity of pipelines to leak. Furthermore, in areas where the vulnerability of the surrounding land is not negligible to leak-induced erosion processes, such zoning can also be interpreted as a partial indicator of the probability of triggering (or initializing) hydrogeological instability phenomena due to water leaks.

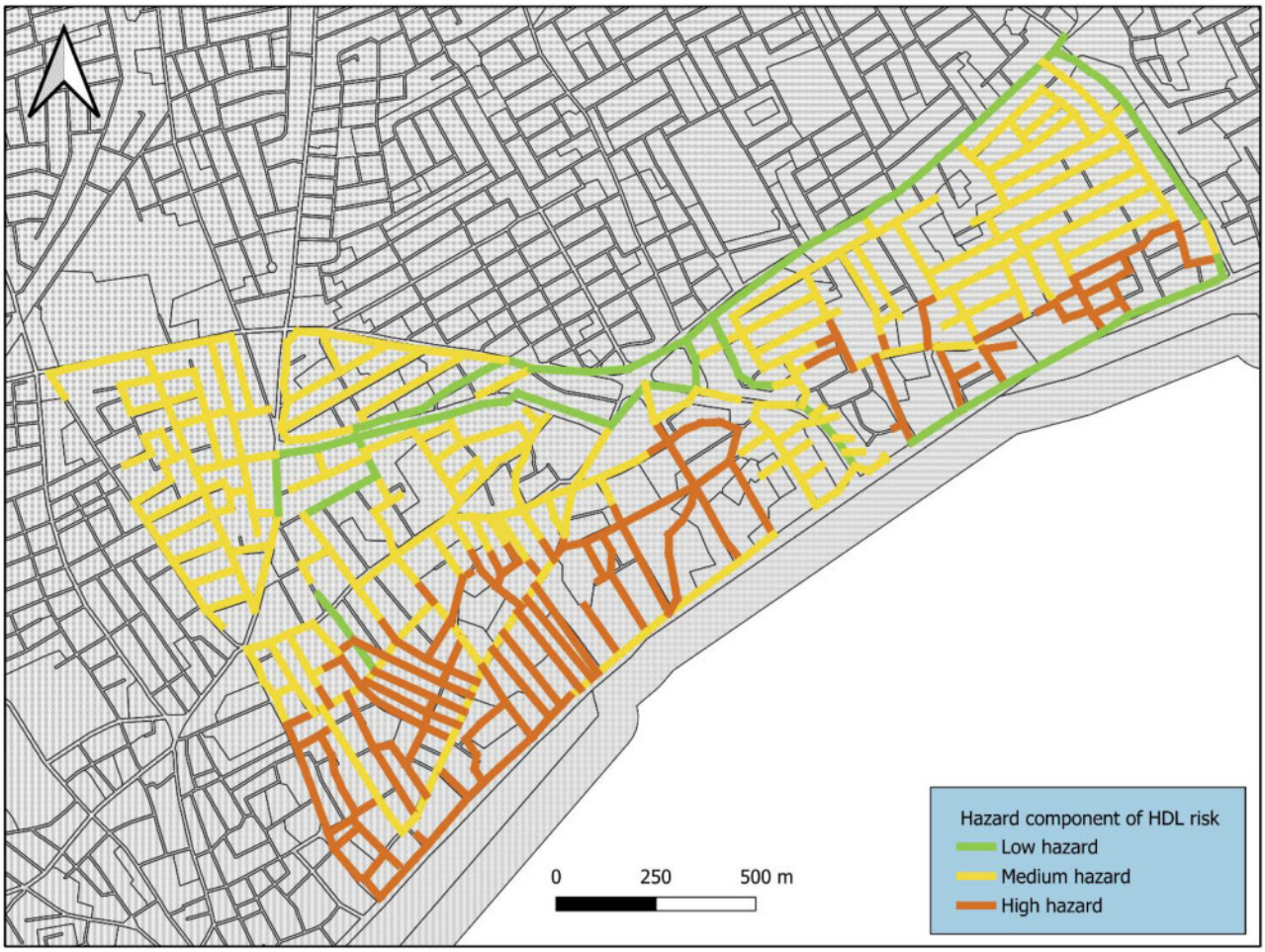

Figure 41. L-Town map of the spatial distribution of the hazard index $H(x)$, obtained from the weighted combination of the structural components $H_D(x)$ (diameter) and operational $H_P(x)$ (pressure).

Risk calculation and final map

The HDL risk map was obtained as a weighted combination of exposure and hazard maps, according to the report:

$$R(x) = \alpha_{\mathrm{E}}\, E(x) \circ \beta_{\mathrm{H}}\, H(x) \tag{26}$$

where $R(x)$ represents the overall risk index in the stretch x; $E(x)$ it is the territorial exposure index, while $H(x)$ it is the hazard index linked to the structural and operational conditions of the network.

In the combination process, greater weight was given to exposure ($\alpha_{\mathrm{E}} > \beta_{\mathrm{H}}$), as it derives from a more complete and articulated analysis, which integrates socio-economic, demographic and infrastructural variables, with respect to the hazard, assessed instead on the basis of a more limited number of components, as the only ones available.

In this way, the sections of the network that have high exposure and a high probability of failure are identified as areas of greater risk.

The resulting HDL risk zoning map (Figure 42) represents a fundamental step for subsequent analyses to optimize the positioning of the sensors, allowing to assign higher priorities to the sections of the

network located in areas characterized by high exposure and hazard, and providing an important information base for the sustainable management of the urban water system.

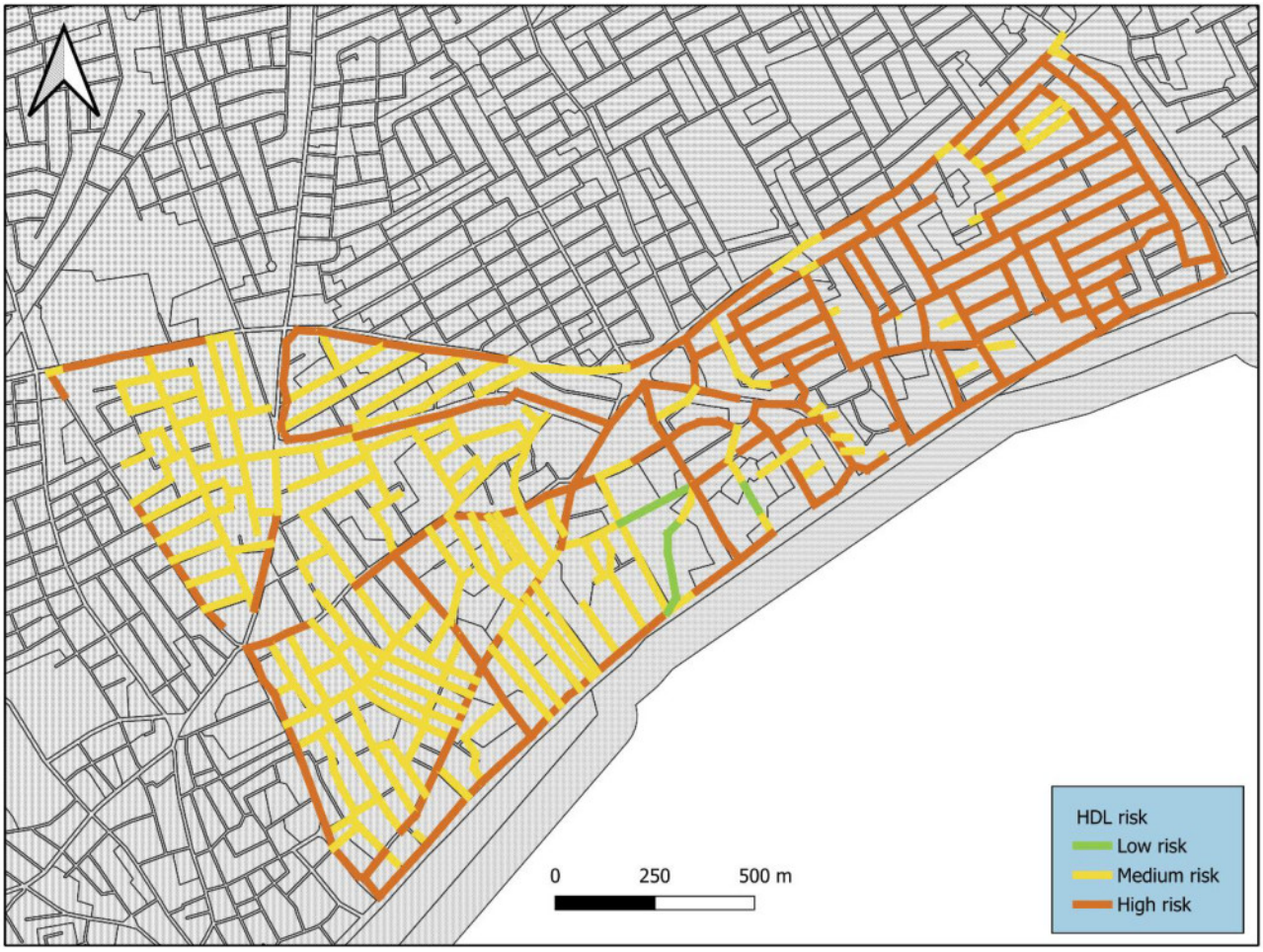

Figure 42. L-Town zoning based on HDL risk. Spatial distribution of the overall risk index $R(x)$, obtained by the weighted combination of the exposure $E(x)$ and hazard $H(x)$ components.

3.2.3. Hydraulic modeling and pressure data generation for L-Town

The hydraulic modeling of the *L-Town* network was carried out through the combined use of EPANET 2.2, the EPyT library and the WNTR (Water Network Tool for Resilience) package. This integration of tools has made it possible to automate the execution of simulations and to have advanced capabilities for leakage management. In particular, WNTR proved to be particularly useful for the modeling of leaks along pipelines, thanks to the availability of already implemented functions that allow the creation of intermediate nodes and the direct assignment of leak parameters (see section 2.3.3.).

Demand modeling

The node demand profile used is the one defined in the *L-Town.inp* file, consistent with what is reported in the benchmark reference model.

As reported in Vrachimis et al. (2022) [120], through a clustering algorithm implemented in QGIS, each node in the network was assigned a set of connected buildings and, consequently, a total built area.

The base demand of each node was calculated as a function of the built area and the composition of use. Three main types of users were considered in the model: residential, commercial and industrial.

Each node has a specific demand pattern for each category, derived from real consumption data and modeled by means of Fourier series, capable of reproducing daily, weekly and seasonal periodicities, as well as the stochastic variability of consumption.

The total demand over time is obtained as a linear combination between the basic demands and their respective time patterns (Figure 43).

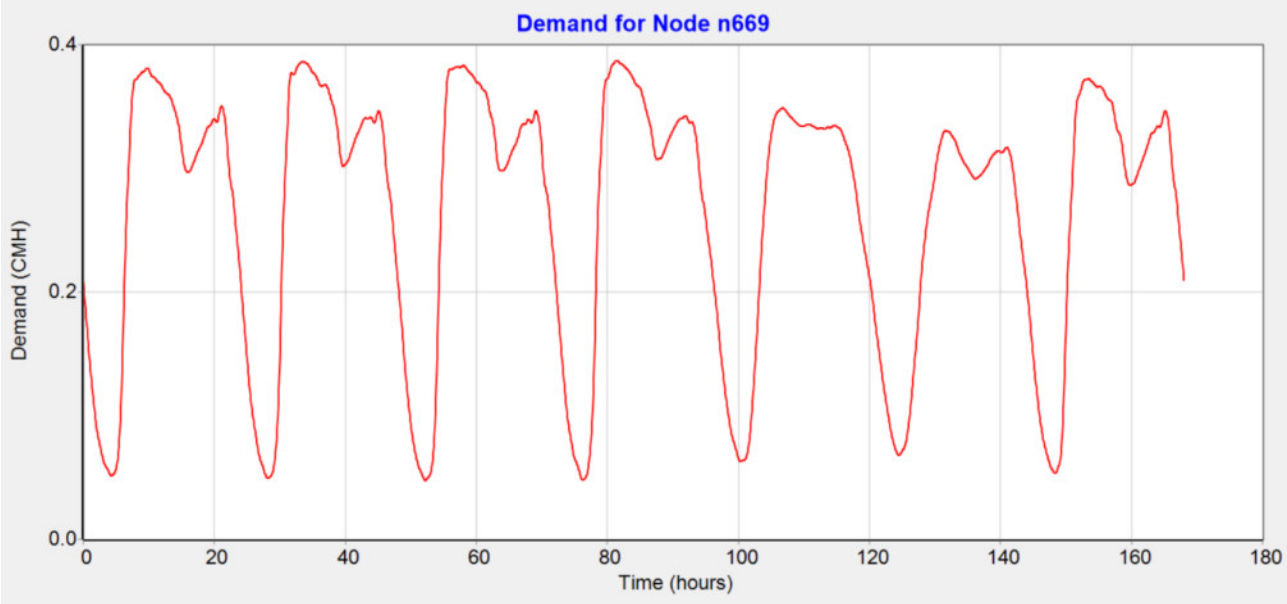

Figure 43. Time trend of water demand at node n669 generated in EPANET. [48]

The peaking factor (DPF), defined as the ratio between the maximum daily demand (MDD) and the average daily demand (ADD), has also been considered in the definition of the demand profiles. For the *L-Town* network, this ratio is between 1.5 and 2.0, in line with what is observed in medium-sized water distribution systems.

Leak modeling

It should be noted, for the reasons mentioned above (see section 3.2.1), that the leak scenarios used in this work are not those provided in the battle but were generated using the WNTR library.

Several hydraulic scenarios were generated, each corresponding to a specific leak condition, with only one active leak at a time. In this case study, leaks were simulated along the pipes and not at the junction nodes, as was the case in the previous case. The total number of scenarios is therefore equal to the number of pipes in the network, since an independent leak was simulated for each pipe.

A *No Leak* reference scenario was also included, which is crucial not only as a term of comparison, but also for the construction of the theoretical signatures used by the localization method.

Each simulation lasted a total of three days, with a constant time step, and the leak was activated at the end of the first day, in order to allow the network to reach steady-state conditions before the trigger. The formulation of the leak is reported in section 2.3.3., which is also referred to to view the specific function dedicated to inserting the leak along a pipeline, within the WNTR library.

In order to ensure a more concise presentation, the results presented in the following sections refer exclusively to leaks with a hole diameter of 2 cm, which is a reasonable size for a localized rupture.

Figure 44 shows an example of the pressure time trend in a representative node of the network, highlighting the pressure drop that occurs at the activation of the leak at the end of the first day of simulation.

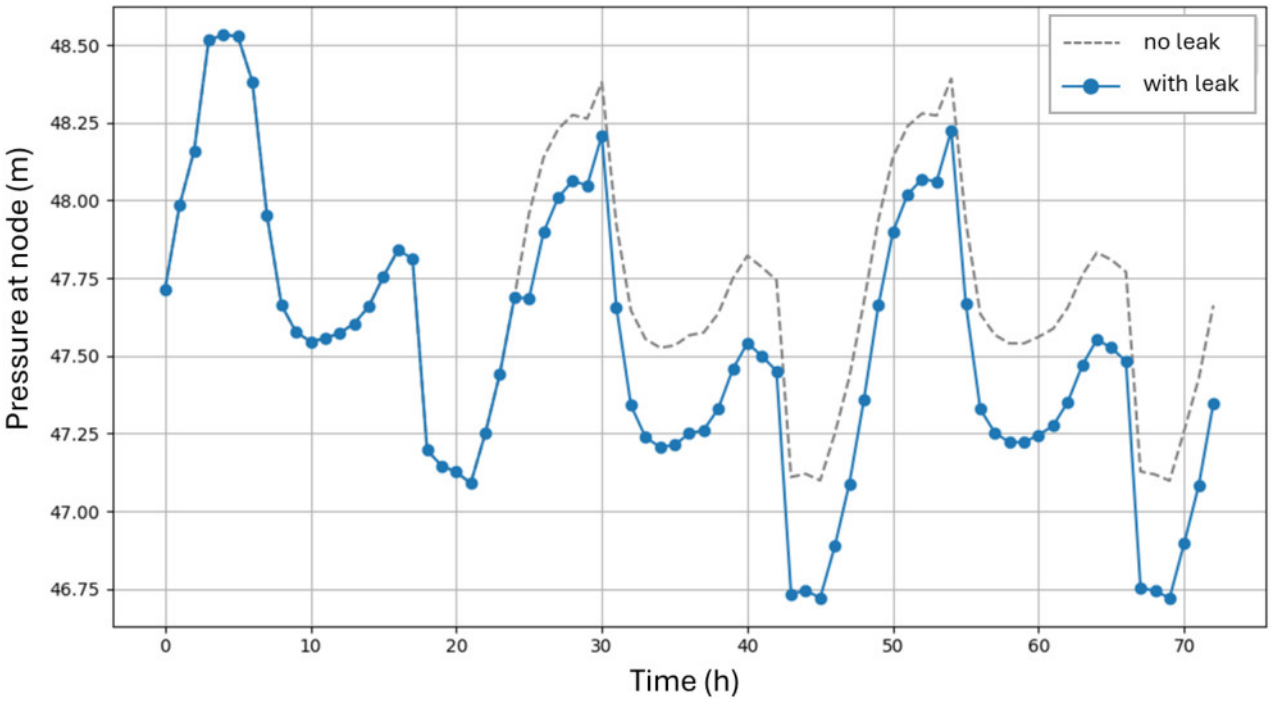

Figure 44. Pressure trend at a representative node of the L-Town network: the pressure drop is observed following the activation of the leak at the end of the first day.

Structure of the generated dataset

The results of all simulations, including the *No Leak* scenario and the various leak scenarios (*Leak Pipe i*), were collected in a single synthetic dataset.

Each row represents a specific moment in time and contains the pressures recorded at the main nodes of the network.

Table 8 reports an extract of the dataset, where two consecutive instants are shown for the *No Leak* scenario and two for a leak scenario (*Leak Pipe i*).

Table 8. Structure of the leak scenario dataset for the L-Town network

Scenario Number	Scenario	Tempo [s]	Node Pressure 1 [m]	...	Node Pressure i [m]	...
0	No Leak	86400	28.442	...	36.095	...
0	No Leak	90000	28.613	...	36.261	...
...	...	...	...	...	...	...
¡	Leak Pipe i	93600	28.790	...	36.436	...
¡	Leak Pipe i	97200	28.971	...	36.617	...
...	...	...	...	...	...	...

This structure allows for a consistent representation of the temporal evolution of pressures for each scenario and a clear distinction between leaky and non-leaky conditions. The resulting dataset forms the input data for the optimization algorithm integrated with the leak localization method described in the next section.

3.2.4. Characteristics and parameter values of the proposed framework for L-Town

The optimization framework adopted for the *L-Town* network is based on the logic described in section 2.5.3. In this case, the optimization process was implemented using the Pymoo library, which provides a consolidated and extensively tested environment for running evolutionary algorithms. The underlying genetic logic remains the same, but the evolutionary operations are handled directly by the Pymoo framework (for further details, see the dedicated section 2.5.3).

Unlike the previous case, the localization of leaks is not performed through a supervised machine learning model, but through an approach based on the sensitivity matrix and the similarity between the signatures that is conducted through the cosine similarity, allowing to identify the most probable location of leak.

Another fundamental innovation concerns the fitness metric, which in this case is expressed as the minimum distance of the hydraulic path between the actual and estimated position of the leak. This formulation, based on distance, allows the algorithm to directly optimize the spatial accuracy of the leak localization, rather than a measure of classification correctness on a percentage basis. Consequently, the goal of optimization is to minimize the value of fitness: a shorter distance indicates a prediction closer to the real location of the leak and therefore a more performing configuration. In other words, it can be said that it is a measure that has a more immediate practical utility.

All the theoretical aspects related to the construction of the sensitivity matrix, the comparison using cosine similarity, and the definition of the minimum hydraulic path were illustrated in Chapter 2, to which the reader is referred for further information. This section instead describes the implementation and parametric aspects used for the application to the *L-Town* case study.

The algorithm begins with the random generation of an initial population of sensor configurations, each represented by a distinct combination of measurement nodes. Each individual corresponds to a possible subset of fixed-size sensors and is constructed in a way that ensures uniqueness within the population.

With each generation, the population evolves through classic genetic operations (selection, crossover, mutation, and elitism) based on the constraints and parameters summarized in Table 9.

Table 9. Optimization constraints and genetic algorithm (GA) parameters for L-Town

Description	Value
Number of genes (sensors)	5
Population size	50 individuals
Crossover probability	90 %
Mutation probability	10 %
Elite percentage	10 %
Number of generations	200

A relatively high mutation rate (10%, typically seen as an acceptable maximum [121]) was deliberately adopted to encourage greater exploration of the solution space, compensating for the limited number of generations. This choice is motivated by the very long simulation times, so fewer generations were preferred but more exploration of the solutions in order to test more parameters with different simulations.

To ensure the robustness of the results, each optimization was repeated several times with different random seeds, then selecting the configuration with the best overall fitness.

Figure 45 shows the trend of fitness function during the evolutionary process. The curve shows a progressive decrease in the fitness value with increasing generations, confirming the algorithm's ability to progressively reduce the hydraulic distance between the actual and estimated loss. After about 75 generations, the trend tends to flatten, indicating the achievement of an optimal or sub-optimal configuration.

It should be noted that the localization of the leaks was evaluated on an independent test dataset, generated with the same algorithmic methods as the one used for the construction of the theoretical signatures. In the test dataset, however, the position of the leak was decentralized with respect to the midpoint of the pipeline (condition on which the basic dataset for the creation of the theoretical signatures was built) in order to simulate different situations but still attributable to the same leak scenarios.

In this type of analysis, accuracy tends to further decrease when additional variations in boundary conditions or data noise are introduced. However, the goal of this work is not to compare the absolute performance of localization methods, nor to refine them, but rather to demonstrate the complete HDL risk mitigation process. The proposed framework therefore highlights the method's ability to guide sensor placement towards configurations that are more functional for the problem being analyzed.

Looking ahead, integrating more advanced localization modules could facilitate the application of this methodology within complex, real-world environments.

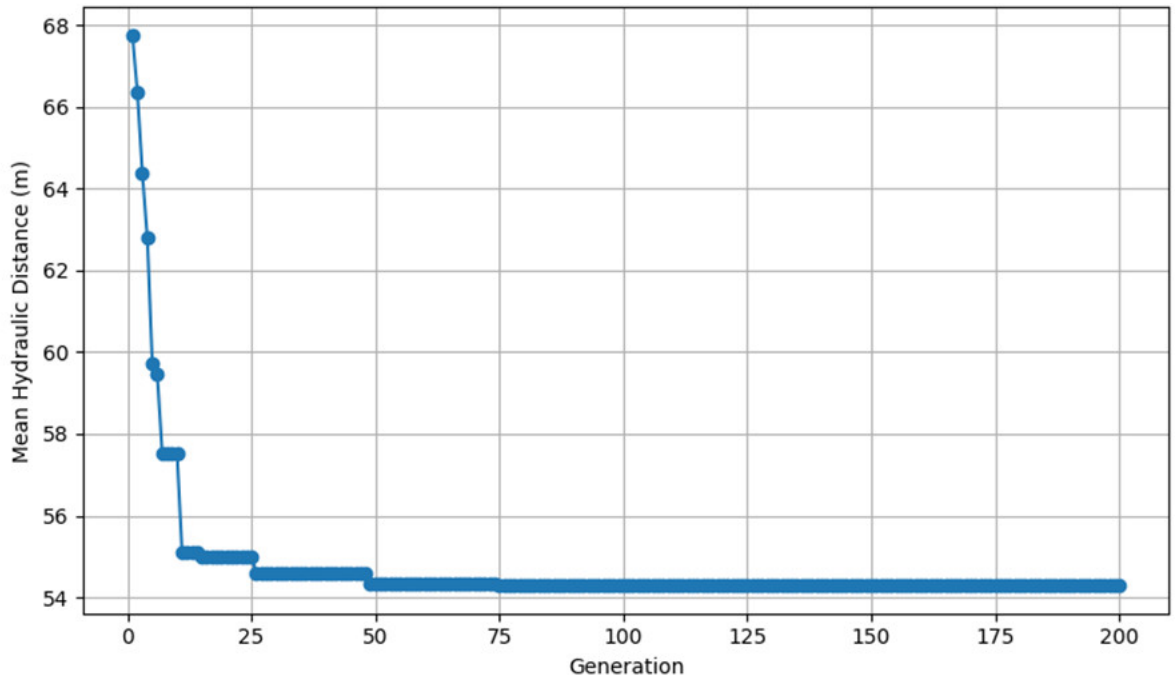

Figure 45. Trend of the fitness function (weighted average hydraulic distance) during the evolutionary process for the configuration with weights W_{R1} = 1, W_{R2} = 2 and W_{R3} = 3. The curve shows a rapid reduction in the average value of the hydraulic distance in the first generations, followed by a stabilization phase around generation 75, indicative of the achievement of an optimal or sub-optimal configuration of the sensors.

3.2.5. Results and discussion for L-Town

As with the previous case study, the method was also tested for the *L-Town* network, considering two different risk configurations: a fictitious zoning, introduced to effectively evaluate the algorithm's ability to guide sensor placement and improve localization accuracy in the most critical areas, and a real zoning, derived from the risk study of the area under consideration, i.e., the zoning obtained using the procedure shown in section 3.2.2.

This dual analysis highlights both the model's flexibility in handling asymmetric and theoretical scenarios, and the feasibility of the solutions when the method is applied to the real risk distribution.

Fictitious zoning

In the first phase of the analysis, a fictitious HDL risk zoning was adopted, with the aim of verifying the ability of the algorithm to direct the positioning of the sensors to increase the accuracy of localization towards high-risk areas.

In this configuration, the zoning has been deliberately defined in a highly asymmetrical way: the highest risk area (R3) has been located on the left side of the network, also including the sector fed by the pump and therefore located at a higher altitude (area C, Figure 30); proceeding to the right we have the medium risk (R2) and low risk (R1) areas.

As in the case of *Real Network 1*, the objective function of optimization has been built by assigning different weights to the various levels of risk, so as to integrate the territorial component in the process of finding the optimal configuration of the sensors.

Before analyzing the graphical results related to the positioning of the sensors, it should be noted that, in the discussion of the maps of the optimal configurations, the real geographical orientation (as shown in Section 3.2.2.) of the network was not maintained. The entire zoning component has in fact been translated into attributes assigned to the relevant sections of the network, and the graphic representation presented here follows the coordinate convention of the reference *.inp* file, in which the network is rotated clockwise with respect to reality, so as to obtain a more extended visualization along the horizontal direction. Therefore, the following maps should be interpreted according to this convention.

The following figures (Figures 46, 47, 48 and 49) show the optimal configurations obtained for each set of weights considered (*uniform*, *progressive*, *reinforced* and *critical*), highlighting the evolution of the distribution of sensors in relation to the different priority levels assigned to the risk areas.

Among the different configurations, it is observed that the arrangement of the sensors in the reinforced (Figure 48) and critical configurations (Figure 49) remains unchanged. This happens because a further increase in weights would lead to an excessive overall imbalance of the system, with a consequent reduction in overall accuracy; This effect is also reflected in the weighted average, further constraining the position of the sensors and limiting the possibility of identifying even more extreme configurations.

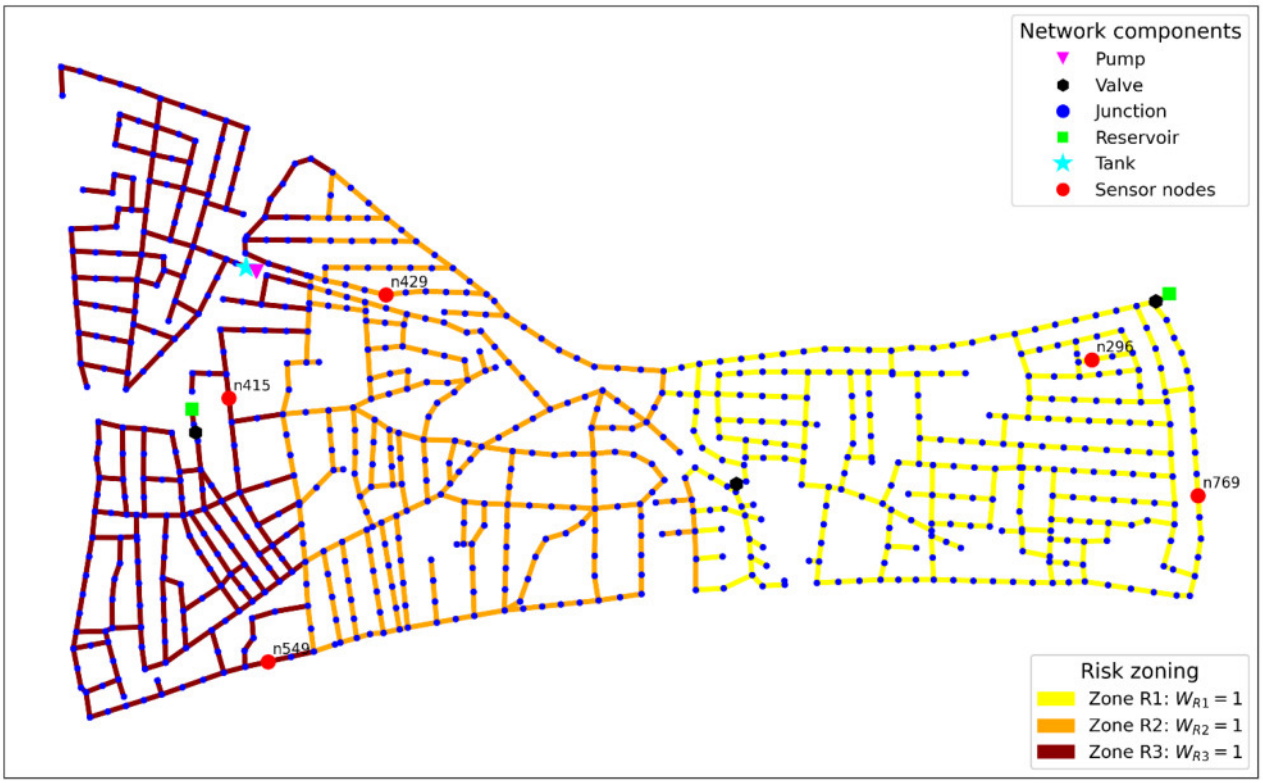

Figure 46. L-Town fictitious zoning – Uniform weight configuration

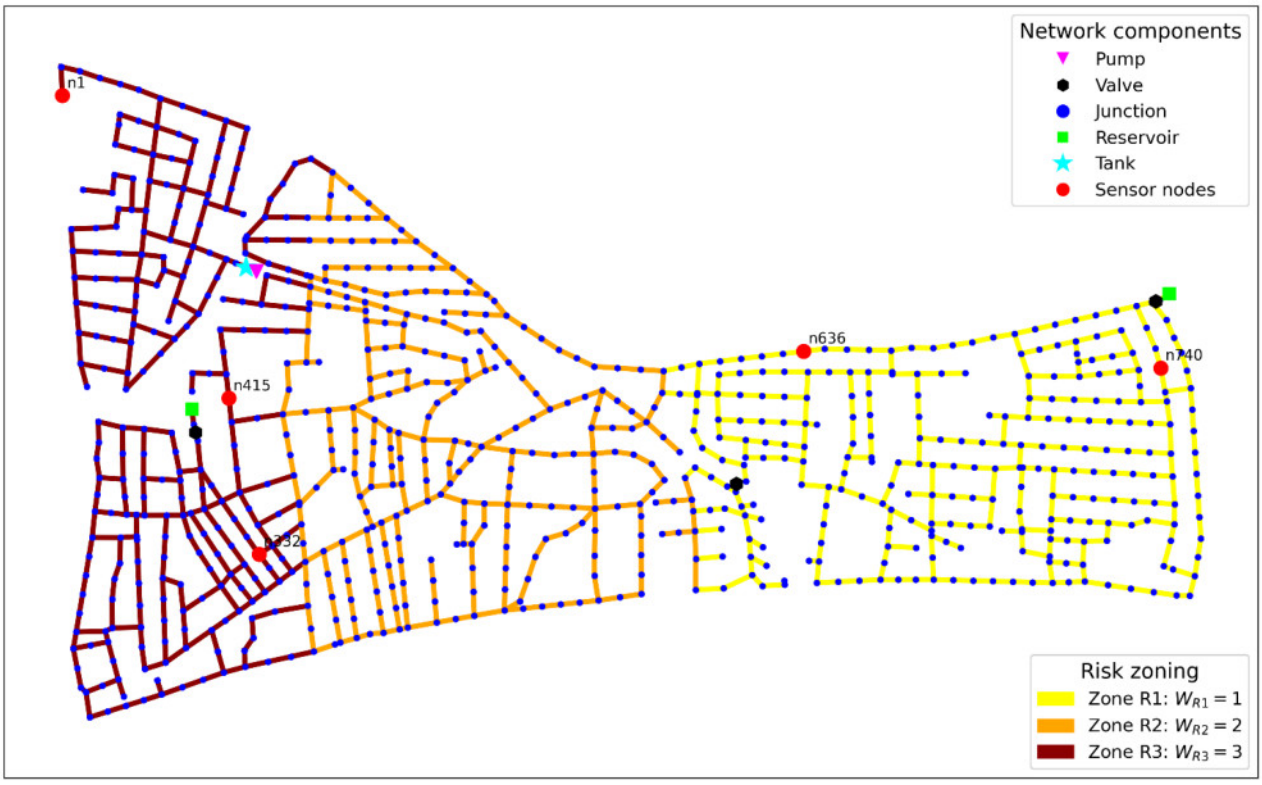

Figure 47. L-Town fictitious zoning – Progressive weight configuration

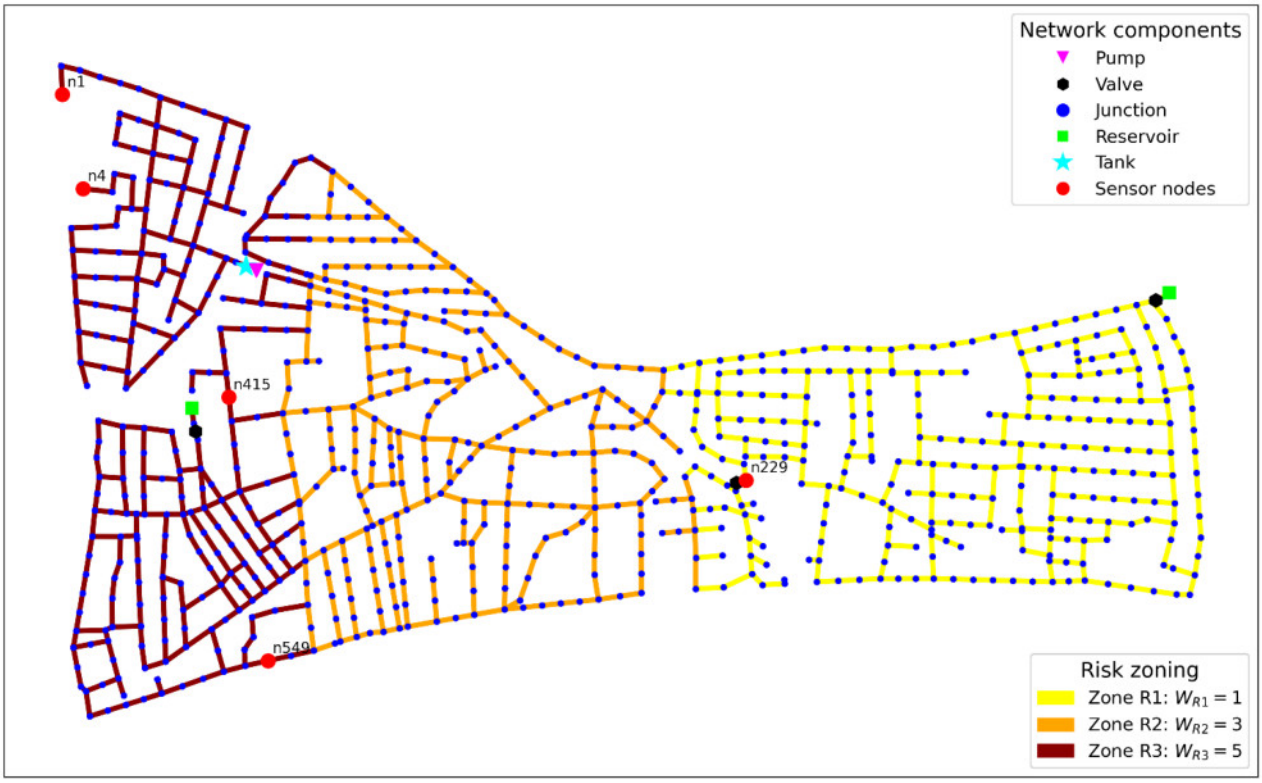

Figure 48. L-Town fictitious zoning – Reinforced weight configuration

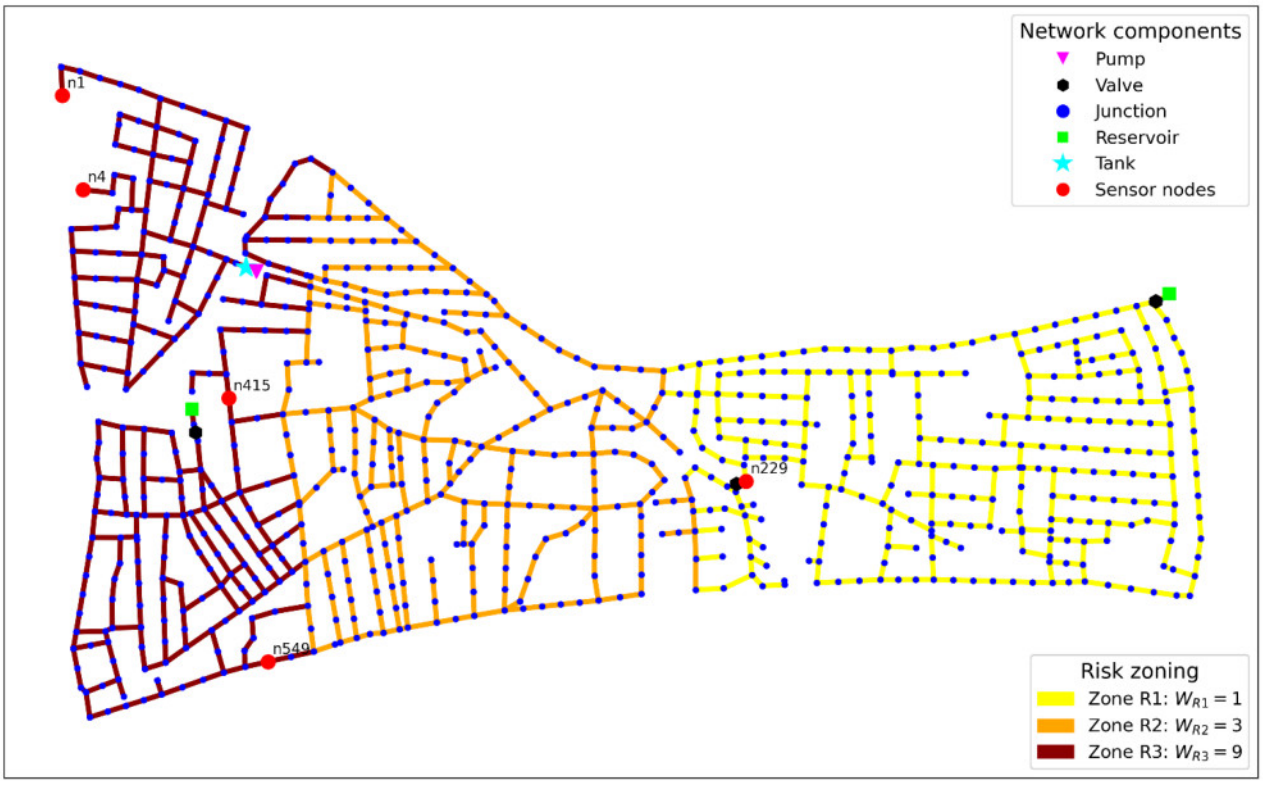

Figure 49. L-Town fictitious zoning – Critical weight configuration

Table 10 summarizes the main numerical results of the optimization, reporting for each configuration:

- overall *fitness* D_w (the global weighted average of the minimum hydraulic distances depending on the zoning);
- the global average of the minimum hydraulic distances calculated in the absence of zoning, D_{GLOBAL};
- and the average minimum hydraulic distances calculated independently for low, medium, and high-risk areas, respectively ($D_{R1,av}$, $D_{R2,av}$, $D_{R3,av}$).

Table 10. Results of the sensor placement optimization process based on the D_w fitness function and HDL-based fictitious risk zoning for different weight configurations in L-Town.

Configuration	W_{R1}, W_{R2}, W_{R3}	*Fitness*, D_w *(m)*	D_{GLOBAL} *(m)*	$D_{R1,av}$ *(m)*	$D_{R2,av}$ *(m)*	$D_{R3,av}$ *(m)*
Uniform	1, 1, 1	46,65	46,65	33,47	30,21	76,44
Progressive	1, 2, 3	54,29	50,18	44,49	37,07	69,21
Reinforced	1, 3, 5	51,73	68,35	120,60	39,04	45,63
Critical	1, 3, 9	49,86	68,35	120,60	39,04	45,63

As the weights assigned to the most critical areas increase, the algorithm tends to favor high-risk regions (R3), progressively reducing the average hydraulic distances in these sectors and improving the ability to locate where the impact of a leak would be most damaging.

This behavior, however, leads to a slight worsening, in terms of accuracy, in intermediate risk areas (R2) and a marked worsening in low-risk areas (R1), accompanied by a decrease in overall accuracy (Figure 50).

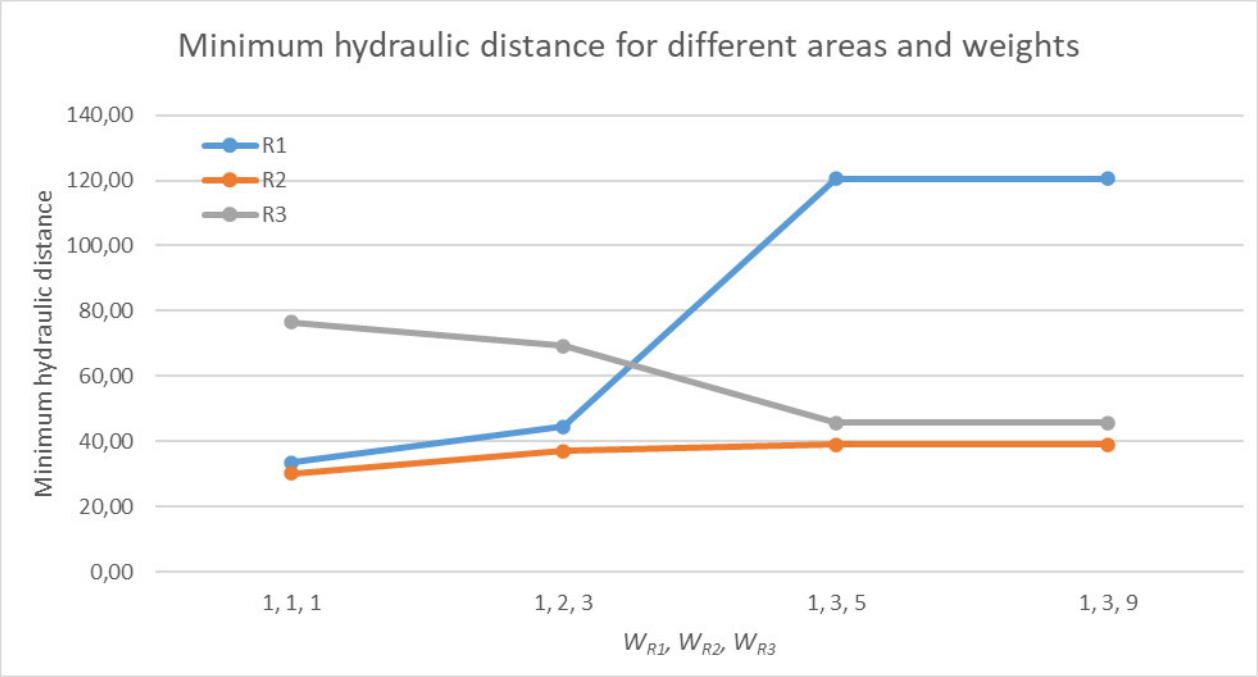

Figure 50. Minimum hydraulic distance for different areas and weights (W_{R1}, W_{R2}, W_{R3}) for L-Town fictitious risk zoning.

As a result, the choice of weights plays a crucial role in finding the right trade-off between improving local performance and maintaining adequate overall accuracy. In this perspective, it is advisable to adopt weights that are not excessively unbalanced towards the most critical areas, so as to ensure a satisfactory balance between local efficiency and overall performance.

Overall, the fictitious zoning confirms the framework's ability to adapt effectively to different weighing conditions, integrating the risk component into the optimization process and orienting the positioning of the sensors in a way that is consistent with the spatial distribution of the risk.

Real zoning

Subsequently, the analysis was extended to the real zoning of HDL risk, elaborated from the data relating to the exposure and hazard components, discussed in Section 3.2.2.

In this case, the areas at high risk are distributed in a less asymmetrical way, even if a more critical area is identified, located on the right side of the network.

Similarly to what was done for the fictitious case, the optimal sensor configurations obtained for the different weight combinations (homogeneous, progressive, reinforced and critical) are shown below (Figures 51, 52, 53 and 54), followed by a summary table of the results in terms of average minimum hydraulic distances calculated for each risk class and for the entire network.

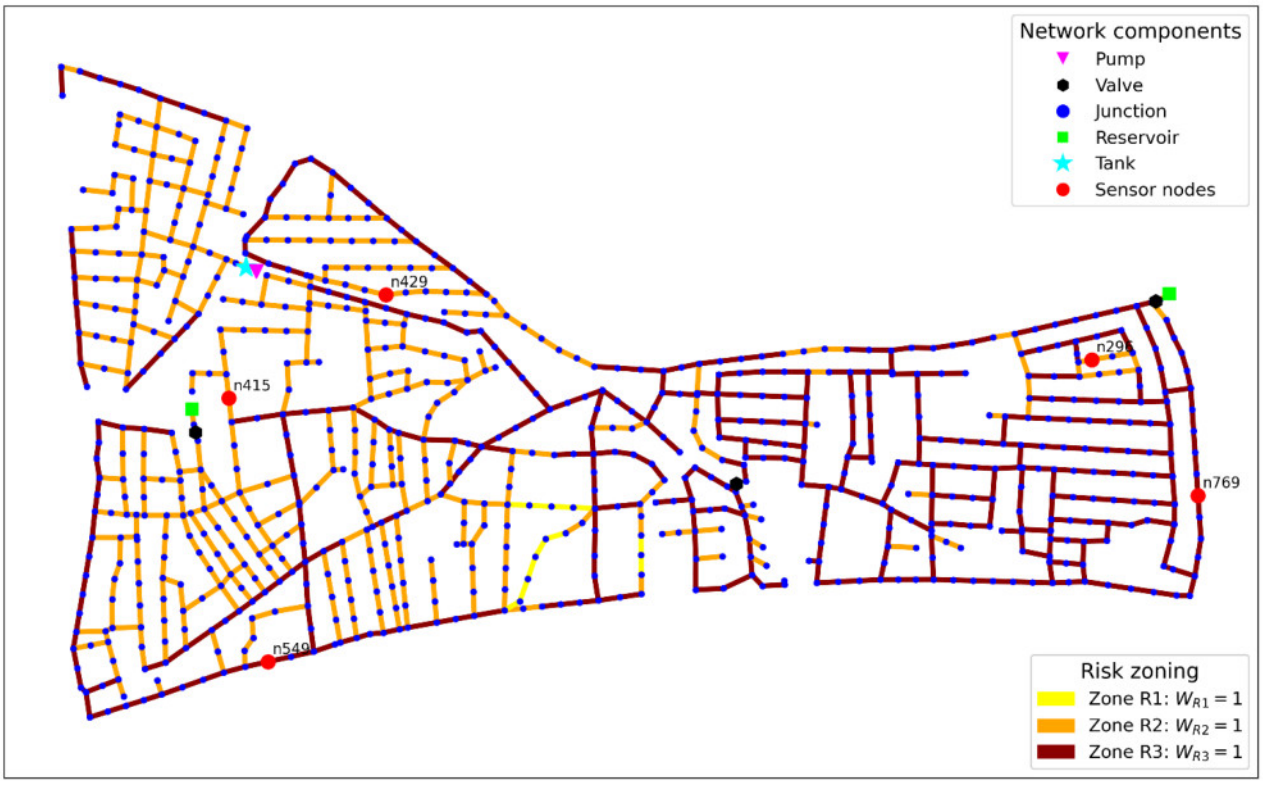

Figure 51. L-Town real zoning – Uniform weight configuration

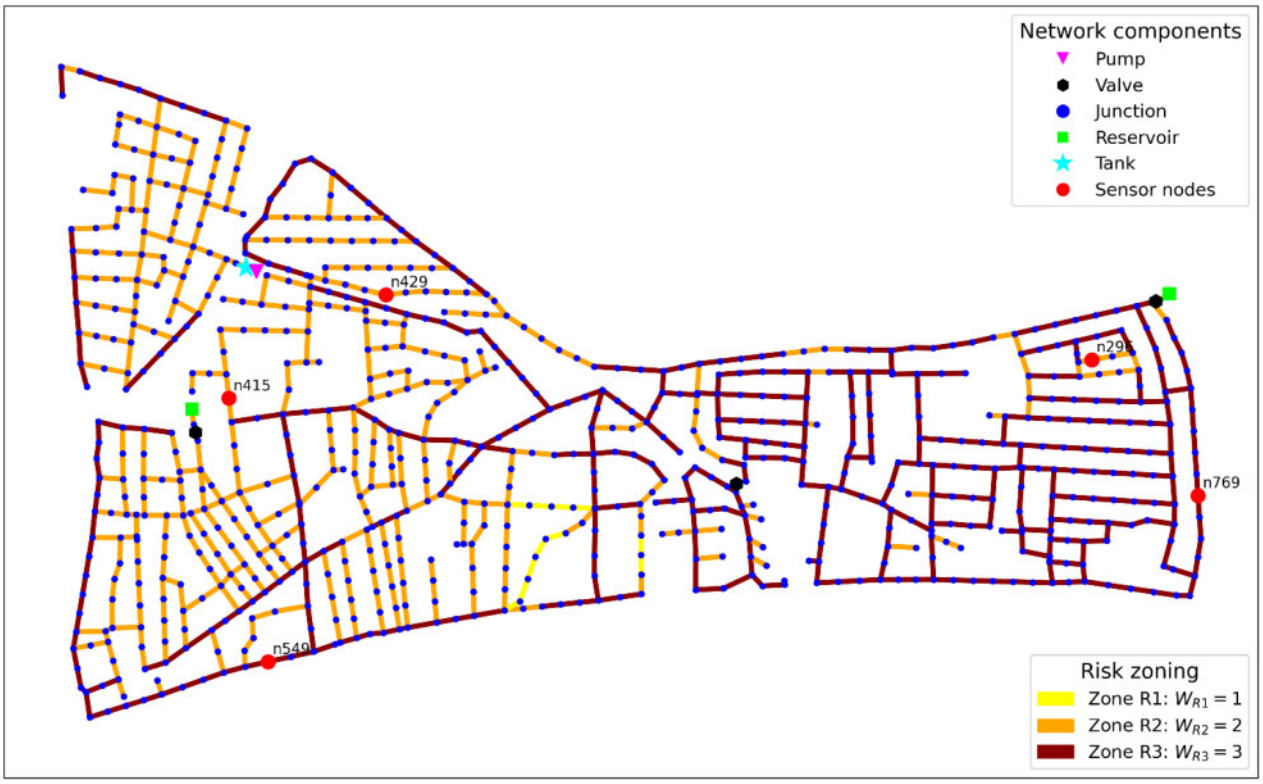

Figure 52. L-Town real zoning – Progressive weight configuration

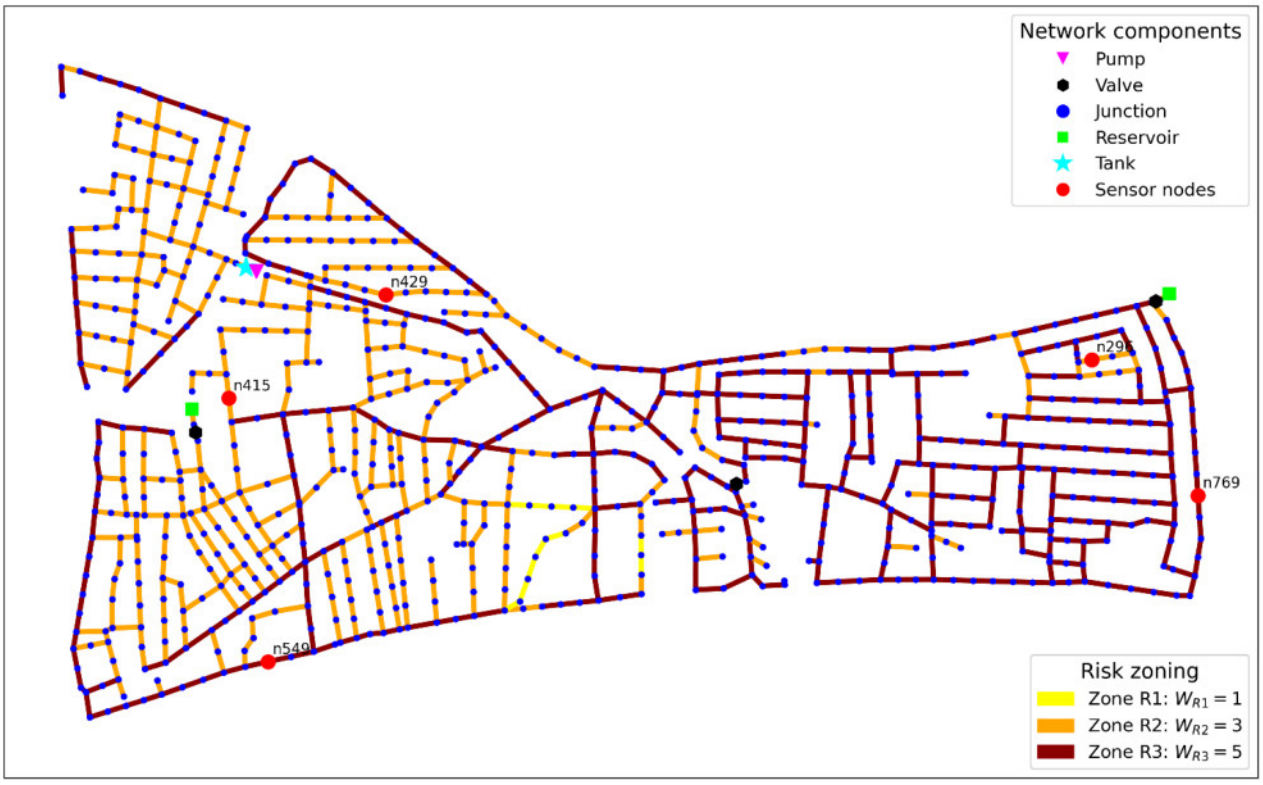

Figure 53. L-Town real zoning – Reinforced weight configuration

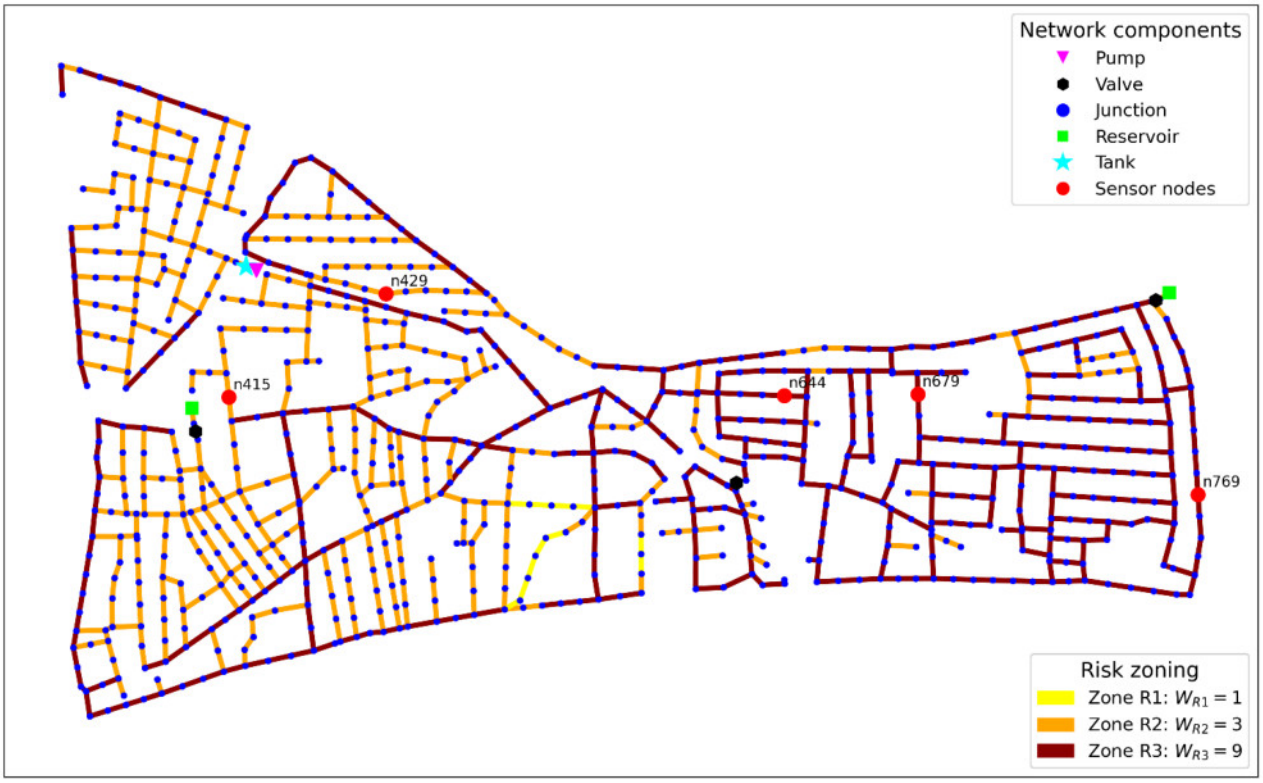

Figure 54. L-Town real zoning – Critical weight configuration

The results relating to the real zoning are summarized in Table 11.

Table 11. Results of the sensor placement optimization process based on the D_w fitness function and HDL-based real risk zoning for different weight configurations in L-Town.

Configuration	W_{R1}, W_{R2}, W_{R3}	Fitness, D_w (m)	D_{GLOBAL} (m)	$D_{R1,av}$ (m)	$D_{R2,av}$ (m)	$D_{R3,av}$ (m)
Uniform	1, 1, 1	46,65	46,65	178,76	53,26	37,04
Progressive	1, 2, 3	44,07	46,65	178,76	53,26	37,04
Reinforced	1, 3, 5	43,42	46,65	178,76	53,26	37,04
Critical	1, 3, 9	49,68	68,35	67,26	66,76	32,74

As can be seen from the data in Table 11, the uniform configuration, i.e. the one in which no zoning is considered (all weights equal), already provides an optimal arrangement of the sensors to maximize the accuracy of location in the areas of higher risk. This indicates that, for the *L-Town* network, the distribution of sensors obtained with uniform weights exhibits a structural layout that is naturally favorable to an effective coverage of critical areas even in the absence of risk weighing.

This behavior is confirmed by observing the results of the *progressive* and *reinforced* configurations: in both cases, there are no improvements in the average minimum hydraulic distances of the areas at higher risk (R3), since further improving the already excellent configuration of the starting sensors is difficult.

This stability shows that the initial arrangement of the sensors is already close to an optimal balance, comparable to the reinforced configuration, capable of guaranteeing excellent levels of accuracy in high-risk areas even with homogeneous weights.

Only in the critical configuration, where the weight assigned to the R3 area is significantly higher, is a partial rearrangement of the sensor layout observed, with two of the five sensors being placed in new positions. This shift leads to an improvement in the average accuracy in the R3 area, but a deterioration in the R2 area and in overall accuracy (Figure 55).

In summary, the results obtained by applying the proposed method to real zoning do not differ significantly from the optimal set-ups deriving from the classic methods of sensor positioning, limiting the necessary changes even in the presence of the risk component. This suggests that, in practical applications, the adoption of the method would not entail high costs of reorganizing the monitoring system, since any movement of the sensors would be limited.

However, such small adaptations can significantly improve leak detection in the long run and enable valuable and timely interventions, particularly in areas with high HDL risk. The latter, if affected by instability phenomena, tend to generate more serious economic and social impacts than other parts of the network, making their monitoring a priority.

Overall, this confirms that, in networks characterized by a substantially balanced distribution of risk, the introduction of zoning does not significantly alter the overall performance, but allows for a more coherent and environmentally conscious arrangement from a territorial and environmental point of view.

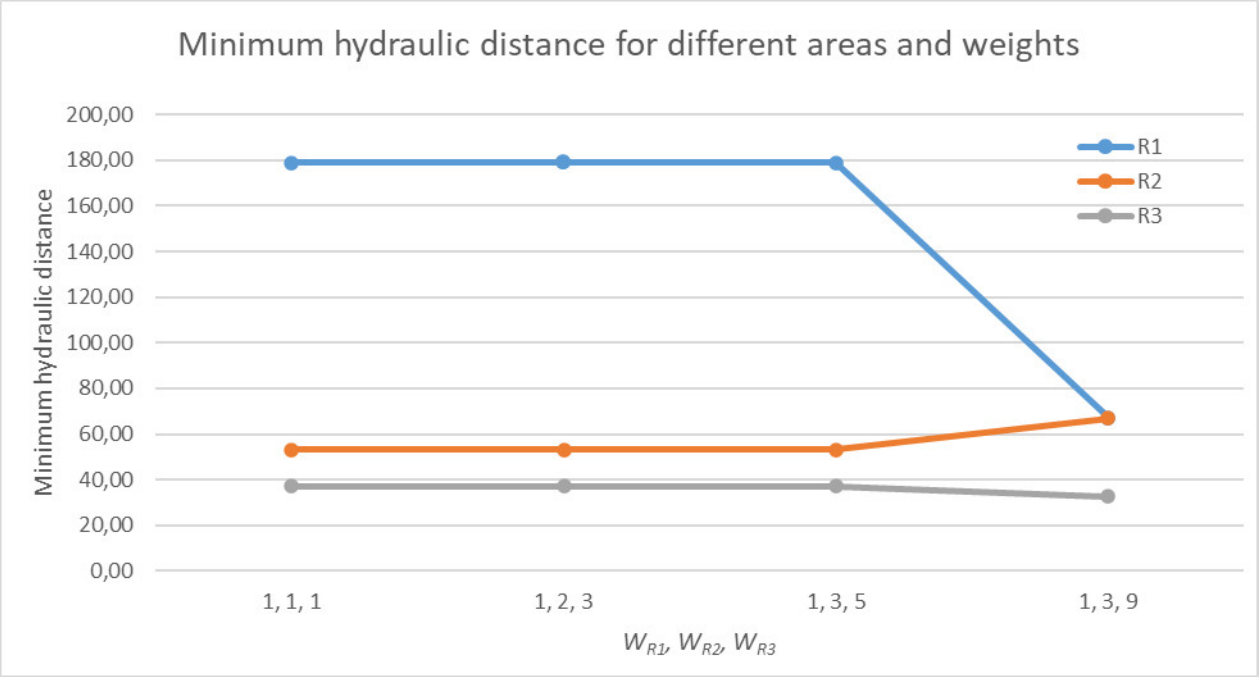

Figure 55. Minimum hydraulic distance for different areas and weights (W_{R1}, W_{R2}, W_{R3}) for L-Town real risk zoning.

.

4. Conclusions

The research work presented addressed the issue of mitigation of the risk of hydrogeological disruption caused by leaks (HDL) in urban water networks in an integrated way, proposing an innovative framework that combines risk zoning, hydraulic modeling, leak location and optimization of sensor positioning. This approach stems from the awareness that leaks are not only a waste of resources, but also a territorial stress factor capable of triggering soil instability phenomena with consequent potential damage to infrastructures, structures, activities and logistics, as well as to the human component.

The developed methodology stands out for its modular nature, which allows the entire process to be considered as a coherent chain of steps: from hydraulic simulation and dataset generation (or data collection), to leak localization and monitoring optimization, aimed at mitigating HDL risk. This systemic approach, which integrates components typically treated separately in literature, represents one of the most innovative aspects of the work. Indeed, while the link between water leaks and hydrogeological instability is widely recognized, integrated operational strategies to reduce their effects in a structured manner are still lacking. This study aimed to fill this gap by providing a conceptual and applicative model capable of uniting land zoning, leak localization, and risk mitigation in a single logical chain. This process optimizes sensor placement to maximize leak localization accuracy in the highest-risk areas, thus laying the foundation for limiting their negative effects through more targeted, timely, and informed maintenance and restoration interventions.

The modularity of the framework also offers a crucial advantage: the possibility of replacing or updating individual modules with more advanced versions or better adapted to the application context. For example, the leak localization module used in this work provided very good results in terms of accuracy; however, these values must be interpreted and contextualized in the light of the method and data used, which are likely to outperform real situations, as they derive from datasets generated with simulators, in which the time series produced and considered present a fairly deterministic behavior and with few uncertainties typical of real systems. Therefore, these results cannot be generalized in a quantitative sense but should be understood as exemplary cases useful to illustrate the entire proposed process. Precisely for this reason, the framework is conceived as an open and adaptable structure, inviting industry operators, researchers and enthusiasts to experiment with more performing modules, such as different leak localization approaches or new accuracy metrics, so as to facilitate its continuous evolution and validation on real cases.

The application of the framework to the two reference networks (*Real Network 1* and *L-Town*) confirmed the flexibility and effectiveness of the proposed method. The results showed that, even in the presence of complex zoning, the framework is able to direct the positioning of the sensors taking into account the overall balance of network monitoring. In particular, in the two cases of real zoning analyzed, the integration of the risk component led to a more balanced distribution of the sensors with respect to the impacts of the leaks on the territory, without however determining significant deviations from the configurations obtained with classical approaches. This results in a reduction of the required operational adjustments, making the method easy to implement and cost-effective from an application standpoint.

This approach is also consistent with the current objectives of increasing urban resilience, favoring more effective and integrated hydrogeological risk management strategies. Overall, the results demonstrate that the proposed framework constitutes a flexible, modular and scalable tool, capable

of supporting the planning of monitoring systems based on territorial priority criteria and in line with modern sustainable development policies.

In real contexts, the application of the framework can represent a valid decision-making support for the proactive management of the territory and water infrastructures, configuring itself as a non-invasive, complementary, replicable and easily implementable method for the optimization of sensor positioning.

Future prospects concern the extension of the method towards multi-objective approaches, able to optimally balance accuracy and costs, and above all the integration with field data from operational networks, in order to further validate its real applicability. The evolution of the framework in this direction will be able to contribute substantially to the construction of smarter, more resilient and sustainable water infrastructures, capable of anticipating and mitigating even the phenomena of hydrogeological instability induced by leaks.

Acknowledgements

At the conclusion of this journey, I would like to thank those who, in various ways, have accompanied the development of this work.

I would like to thank my supervisor Prof. Renata Della Morte, the doctoral program coordinator Prof. Antonio Occhiuzzi, and my co-supervisors Prof. Luca Cozzolino and Dr. Giada Varra, for providing the institutional framework necessary for my doctoral studies and for the administrative support that made the finalization of this research possible.

A heartfelt recognition goes to Andrea Cominola for welcoming me during my research period abroad at TU Berlin. In just three months of collaboration, I received the critical and technical tools that were essential in making my research more rigorous and innovative, representing undoubtedly one of the most valuable contributions of my entire PhD.

An affectionate thought goes to my colleagues in the legendary 514 South office. You are all exquisite people to whom I sincerely wish the very best; we could certainly have shared one more pizza together between one deadline and the next, but I am sure there will be many more opportunities to do so.

Finally, my most intimate thanks go to my family. You have been my safe harbor with your simplicity and your constant hope and trust. Even in moments of profound difficulty and exhaustion, which I chose to face in silence to avoid burdening you with my worries, your closeness was the silent strength that allowed me to keep moving forward. Reaching the end of this path today with an innovative topic, developed with great autonomy and of which I am proud, is a milestone I deeply share with you, in the hope that it represents only the first step toward new and inspiring future developments.

Dissemination of Thesis Research Findings

Medio, G., Varra, G., İnan, Ç.A., Cozzolino, L., Della Morte, R.: *Sinkhole Risk-Based Sensor Placement for Leakage Localization in Water Distribution Networks with a Data-Driven Approach*. Sustainability, 16, 5246 (2024). https://doi.org/10.3390/su16125246

Medio, G.: *Mitigation of Hydrogeological Risk from Water Leakage in Urban Water Distribution Networks using Data-Driven Approach*, PhD Days and Marchi Lecture, Trieste 27–28 June 2024. Presentation.

Medio, G., Varra, G., İnan, Ç.A., Cozzolino, L.: *Mitigazione del Rischio Idrogeologico da Perdite nelle Reti di Distribuzione Idrica Urbane mediante Algoritmi Genetici e Tecniche di Machine Learning*, XXXIX Convegno Nazionale di Idraulica e Costruzioni Idrauliche, Parma 15–18 September 2024. Poster. https://doi.org/10.5281/zenodo.13584918

Medio, G., Varra, G., Della Morte, R., Cozzolino, L.: *Leakage Localization in Water Distribution Networks for the Minimization of Hydrogeological Disruption: A Machine Learning Approach*, CSDU Days 24, 2–4 December 2024, Università della Calabria. Presentation.

Medio, G., Cozzolino, L., Varra, G., Della Morte, R., Cominola, A.: *Mitigation of Hydrogeological Risk Caused by Leakage in Urban Water Distribution Networks: An Optimal Sensor Placement Approach*. EGU General Assembly 2026, Vienna, Austria & Online | 3–8 May 2026. Abstract accepted.

Research Period Abroad

Technische Universität Berlin (TU Berlin), Germany;

17 February 2025 – 16 May 2025;

Host Group: *Smart Water Networks*;

This research period was instrumental to the development of the PhD thesis, providing the necessary framework to advance the core methodology. The activities were directly integrated into the doctoral research as follows:

- Methodological Advancements: Testing of methods based on the sensitivity matrix and refinement of evaluation metrics through the minimum hydraulic path.
- Software & Optimization: Implementation of the WNTR (Water Network Tool for Resilience) library and comparative analysis of advanced genetic algorithms, with the final selection of Pymoo.

Additional scientific output by Gabriele Medio on non-thesis related topics

Medio, G., Severino, V., Teta, R., Endreny, T., Lega, M.: *Hierarchical Monitoring of Water Quality: Coordinating the Spatiotemporal Resolution of Multilayer and Multispectral Sensors to Characterize Pollution*. Presented at the Waste Management and Environmental Impact, 2022 August 16 (2022). doi:10.2495/WMEI220011 WMEI22001FU1.pdf

Lega, M., Medio, G., Endreny, T., Esposito, G., Costantino, V., Teta, R.: *Attribution of Pollution Discharges in Coastal Waters During the Covid-19 Lockdown Using Remote Sensing and Bioindicators*. Presented at the waste management and environmental impact 2022 August 16 (2022). doi:10.2495/WMEI220151. WMEI22015FU1.pdf

Lega, M., Medio, G., Endreny, T., Casazza, M., Esposito, G., Costantino, V., Teta, R.: *Cyanobacterial Biomonitoring in Lake Avernus During the COVID-19 Pandemic: Integrating Remote Sensing and Field Data for Pollution Source Detection*. IJCMEM. 11, 135–141 (2023). https://doi.org/10.18280/ijcmem.110301

Esposito, G., Glukhov, E., Gerwick, W.H., Medio, G., Teta, R., Lega, M., Costantino, V.: *Lake Avernus Has Turned Red: Bioindicator Monitoring Unveils the Secrets of "Gates of Hades"*. Toxins. 15, 698 (2023). https://doi.org/10.3390/toxins15120698

Varra, G., Medio, G., Cozzolino, L., Della Morte, R., Tartaglia, M., Fiduccia, A., Agostino, I., Zammuto, A.: *Understanding the Impacts of Extreme Hydro-Meteorological Events on Railway Infrastructure*. Giornate dell'idrologia della Società Idrologica Italiana 2023, Matera, 13-15 September 2023. Poster. https://doi.org/10.5281/zenodo.10200499

Di Matteo, V., Esposito, G., Glukhov, E., Gerwick, WH., Medio, G., Teta, R., Lega, M., Costantino, V.: *Lake Avernus has Turned Red: Bioindicator Monitoring unveils the Secrets of "Gates of Hades"*. Presented at the 2nd Workshop SETAC Italian Language Branch, October 11 (2023). 2nd_workshop SETAC ILB_AbstractBook.pdf

Lega, M., Medio, G., Severino, V., Casazza, M., Endreny, T., Teta, R.: *Coastal Water Pollution Characterization: Enhanced Situational Awareness through Multiscale Data Acquisition and Analysis*. IJEI. 7, 133–140 (2024). https://doi.org/10.18280/ijei.070115

Varra, G., Medio, G., Cozzolino, L., Della Morte, R.: *Flood Impact Simulation of Dam-Break Events with a Binary Porosity based Shallow Water Model: The Case of Tous Dam in Spain*. XXXIX Convegno Nazionale di Idraulica e Costruzioni Idrauliche, Parma 15–18 September 2024. Presentation. https://doi.org/10.5281/zenodo.13584918

Notation

A	The equivalent area of the hole (m^2)
ADD	Average Daily Demand
$A_{E1,i}$, $A_{E2,i}$, $A_{E3,i}$	Localization accuracy values for the nodes belonging to low, medium, and high exposure areas, respectively
$A_{E2,av}$, $A_{E3,av}$	Average local accuracy in medium and high exposure areas, respectively
AE–RF	Autoencoder–Random Forest
A_{GLOBAL}	Global average localization accuracy across the entire network
AI	Artificial Intelligence
AMR	Automatic Meter Reading
ANN	Artificial Neural Networks
$A_{p,max}$	Maximum weighted accuracy per generation of the genetic algorithm
API	Application Programming Interface
$A_{surf,i}$	Surface area of the i-th census section
A_w	Weighted Accuracy
$A_{z,i}$	Localization accuracies for nodes belonging to risk class *z*
BattLeDIM	Battle of the Leakage Detection and Isolation Methods
BR	Bedrock
BRFs	Break Rate Functions
BT2021	2021 Territorial Bases according to ISTAT
BTs	Territorial Bases
C	Hazen-Williams roughness coefficient
c_1	Positive constant that represents the uniform combination of hazards and vulnerabilities ($H_0 \cdot V_0$)
CC BY-SA 4.0	Creative Commons Attribution-ShareAlike 4.0 International License
C_d	The exhaust coefficient (dimensionless), typically equal to 0.75 in turbulent conditions
CFD	Computational Fluid Dynamics
CFS	Cubic Feet per Second
CLMS	Copernicus Land Monitoring Service
COF	Consequence of Failure
COIN	Computational Optimization and Innovation Laboratory
CPIS	Cumulative Pressure-Induced Stress
CyStat	Statistical Service of Cyprus
D	Pipe diameter
DC(t)	Demand coefficient at time t
DEM	Discrete Element Method
D_{Global}	The global average of the minimum hydraulic distances calculated in the absence of zoning
$D_{hydraulic}$	Minimum hydraulic distance between the midpoints of two network pipes
D_i	The demand assigned to the node i
DIC	Digital Image Correlation

d_j	Actual demand delivered at node j based on the available pressure head P_j
D_j	The full normal demand at node j
DL	Deep Learning
DMA	District Metered Area
DPF	Demand peaking factor
$D_{R1,av}$, $D_{R2,av}$, $D_{R3,av}$	The average minimum hydraulic distances calculated independently for low, medium, and high-risk areas, respectively
DT	Decision Tree
D_w	The global weighted average of the minimum hydraulic distances depending on the zoning
$D_{z,i}$	Minimum hydraulic distance for the scenario *i* in the risk zone *z*.
E	Exposure component
E(x)	The exposure index in the stretch x
E1, E2, E3	Low, medium, and high levels of exposure respectively
$E_{ba}(x)$	The exposed qualitative value of built assets
EC_i	Emission coefficient at node i
EEA	European Environment Agency
EFTA	European Free Trade Association
EPA	Environmental Protection Agency
EPANET	Environmental Protection Agency Network Evaluation Tool
EPANET-CPA	EPANET – Cyber-Physical Attacks
$E_{pd}(x)$	The exposure in terms of population density
EPR	Evolutionary Polynomial Regression
EPSG	European Petroleum Survey Group – The world standard for identifying coordinate reference systems.
EPyT	EPANET-Python Toolkit
$E_{rs}(x)$	The importance of the road system
EU	European Union
f	Friction factor (dependent on ϵ, *d*, and *q*)
FDM	Finite Difference Method
FUAs	Functional Urban Areas
G	The acceleration of gravity (m/s^2)
GA	Genetic algorithm
GIS	Geographical Information System
GPR	Geophones, Penetration Radar
GRA	Grey Relation Analysis
H	Hazard
H	The manometric pressure expressed as piezometric height (m)
H(x)	The hazard index in the stretch x
$h_{act,i}$	Non-standardized pressure head
$H_D(x)$	Hazard component related to pipe diameter
HDL	Hydrogeological disruption caused by leaks
$h_f(Q)$	Represents the pressure drop distributed along the pipeline
H_i	The piezometric height at the end i of the pipe

h_i	Piezometric head at node i
H_j	The hydraulic head at node j
H_k	The piezometric height at the end k of the pipe
h_m	The localized loss associated with bends, valves, or fittings
$H_P(x)$	Hazard component related to pipeline operating pressure
HTTP	HyperText Transfer Protocol
$h_{st,i}$	Standardized pressure head
ILP	Identify, Localize, Pinpoint
inhab.	Inhabitants
InSAR	Interferometric Synthetic Aperture Radar
IoT	Internet of Things
ISTAT	Italian National Statistics Institute
KMZ	Keyhole Markup Language Zipped
L	Pipe length (ft);
LDA	Linear Discriminant Analysis
LLS	Leakage Location System
LS	Limestone
L_{leak} *and* L_{pred}	Lengths of the actual and predicted pipes affected by leakage, respectively
Matplotlib	Python library for creating static, animated, and interactive visualizations
MaxEnt	Maximum Entropy
MDD	Maximum Daily Demand
min_path	WNTR function to calculate the distance along the minimum hydraulic path
$M_k(P)$	Localization accuracy of the node cluster Ω_k
ML	Machine Learning
MPI	Maximum Pressure Indicator
N	Manning roughness coefficient
$N_{Ak}(P)$	Number of correct predictions relative to the cluster Ω_k
N_k	The total number of leak scenarios generated by nodes in the cluster Ω_k
NNET	Neural Network
NRW	Non-Revenue Water
NTC 2018	Norme Tecniche per le Costruzioni - Technical Standards for Construction
NumPy	Numerical Python
N_z	Number of leak scenarios in risk zone *z*
N_{E1}, N_{E2}, N_{E3}	Number of nodes potentially subject to leak in the low, medium and high exposure areas respectively
N_{leak} *and* N_{pred}	Respectively, the pairs of end nodes of the pipes affected by the actual and predicted leaks
Ø	Diameter
OMI	Osservatorio Mercato Immobiliare – Real Estate Market Observatory
OSP	Optimal Sensor Placement
P_0	Minimum pressure threshold
Pandas	Python library for high-performance data manipulation via DataFrames.
PCA	Principal Component Analysis
PDD	Pressure Dependent Demand

PD_i	Resident population density of the *i*-th census section
P_f	Full demand pressure
PhD	Philosophiae Doctor
P_j	The available pressure head at node j
P_{min}	The threshold pressure below which the delivered demand is zero
POF	Probability of Failure
Pop_i	Resident population per i-th census section
PRV	Pressure Reducing Valve
P_{ser}	The service limit pressure for node j
Pymoo	Python multi-objective optimization framework.
q	Flow rate
q_{Bi}	Basic demand at node i
QGIS	Quantum Geographic Information System - Open-source GIS software
Q_i	Simulated leak flow rate by emitters placed at nodes i
$q_i(t)$	Demand at node i at time t
Q_j	The flow rate in the pipeline j
Q_L	Loss rate (m^3/s)
R	Risk
r	Residual vector of pressure deviations between field measurements and model simulations
$R(x)$	The risk index in the stretch x
R1, R2, R3	Low, medium, and high levels of risk respectively
RBI	Registro di Base degli Individui - Basic Register of Individuals
RCAs	Road Collapse Accidents
ResNet	Residual Network
RF–RFE	Random Forest - Recursive Feature Elimination
RSBL	Registro Statistico di Base dei Luoghi - Basic Statistical Register of Places
S	Sensitivity matrix
$\bar{S}$	Normalized sensitivity matrix
SC	Sandy Clay
SCADA	Supervisory Control and Data Acquisition
scikit-learn library	Open-source library used for Machine Learning in Python
SciPy	Python library for advanced mathematical, scientific, and engineering computing
SEDP	Soil Erosion due to Defective Pipes
semi-JMI	Semi-supervised Joint Mutual Information
SIT	Sistema Informativo Territoriale - Territorial information system
SMAT	Società Metropolitana Acque Torino -
s_N	Sensitivity vector of theoretical pressure drops associated with the N-th leak scenario
SPC	Statistical Process Control
SRI	Sinkhole Risk Index
SVM	Support Vector Machines
SWL	Soil and Water Loss
SWN	Smart Water Networks

TDE	Topological Differential Evolution
UGC	Urban Ground Collapse
Urban Atlas 2018	High-resolution cartographic dataset produced by CLMS and coordinated by EEA
UTF-8	Unicode Transformation Format, 8-bit
UTM	Universal Transverse Mercator - Map projection system
V	Vulnerability component
w_{ba}, w_{pd}, w_{rs}	The weights assigned to the components $E_{ba}(x)$, $E_{pd}(x)$ and $E_{rs}(x)$ respectively
WCS	Web Coverage Service
w_D, w_P	The weights assigned to the components $H_D(x)$ and $H_P(x)$ respectively
WDN	Water Distribution Networks
WFS	Web Feature Service
WGS	World Geodetic System
WMS	Web Map Service
WNTR	Water Network Tool for Resilience
W_{R1}, W_{R2}, W_{R3}	Weights associated with low, medium and high risk areas respectively
WSNs	Wireless Sensor Networks
W_z	Weight parameter associated with the risk zone z
W_{E1}, W_{E2}, W_{E3}	Weights associated with low, medium and high exposure areas respectively
x_1, x_2	Two generic points in the spatial domain
z	Risk zone
Z_j	The elevation of node j
α	Weight of the exposure component
α_E, β_H	The weights assigned to the exposure index E(x) and the hazard index H(x) respectively
β	Weight of the hazard component
δ	Pressure exponent
ϵ	Darcy-Weisbach roughness coefficient (ft)
$\mu_{DC}(t)$	Average demand coefficient
μ_{hi}	Mean of non-standardized pressures head
σ_{hi}	Standard deviation of non-standardized pressures head
Ω_k	Localization cluster consisting of nodes
.inp	Standard input file for EPANET water network simulations
$\partial p/\partial f$	The change in pressure at the node due to a change in flow rate (or leak)
$\circ$	Generic combination operation between components

References

1. Liemberger, R., Wyatt, A.: Quantifying the global non-revenue water problem. Water Supply. 19, 831–837 (2019). https://doi.org/10.2166/ws.2018.129
2. Kingdom, B., Liemberger, R., Marin, P.: The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries - How the Private Sector Can Help: A Look at Performance-Based Service Contracting. The World Bank, Washington, DC (2006)
3. Evaristo, J., Jameel, Y., Tortajada, C., Wang, R.Y., Horne, J., Neukrug, H., David, C.P., Fasnacht, A.M., Ziegler, A.D., Biswas, A.: Water woes: the institutional challenges in achieving SDG 6. Sustain. Earth Rev. 6, 13 (2023). https://doi.org/10.1186/s42055-023-00067-2
4. AVK Group: AVK Non-Revenue Water Solutions. (2017)
5. Istituto Nazionale di Statistica (ISTAT): Le statistiche dell'Istat sull'acqua. Anni 2020–2022. (2023)
6. Becher, O., Smilovic, M., Verschuur, J., Pant, R., Tramberend, S., Hall, J.: The challenge of closing the climate adaptation gap for water supply utilities. Commun. Earth Environ. 5, 356 (2024). https://doi.org/10.1038/s43247-024-01272-3
7. He, C., Liu, Z., Wu, J., Pan, X., Fang, Z., Li, J., Bryan, B.A.: Future global urban water scarcity and potential solutions. Nat. Commun. 12, 4667 (2021). https://doi.org/10.1038/s41467-021-25026-3
8. Guo, J., Zhang, Y., Li, Y., Zhang, X., Zheng, J., Shi, H., Zhang, Q., Chen, Z., Ma, Y.: Model experimental study on the mechanism of collapse induced by leakage of underground pipeline. Sci. Rep. 14, 17717 (2024). https://doi.org/10.1038/s41598-024-68824-7
9. D'Aniello, A., Cimorelli, L., Pianese, D.: Leaking pipes and the urban karst: a pipe scale numerical investigation on water leaks flow paths in the subsurface. J. Hydrol. 603, 126847 (2021). https://doi.org/10.1016/j.jhydrol.2021.126847
10. Ali, H., Choi, J.: Data on manmade sinkholes due to leakage in underground pipelines in different subsurface soil profiles. Data Brief. 34, 106740 (2021). https://doi.org/10.1016/j.dib.2021.106740
11. Wang, X.-W., Xu, Y.-S.: Investigation on the phenomena and influence factors of urban ground collapse in China. Nat. Hazards. 113, 1–33 (2022). https://doi.org/10.1007/s11069-022-05304-z
12. Karoui, T., Jeong, S.-Y., Jeong, Y.-H., Kim, D.-S.: Experimental Study of Ground Subsidence Mechanism Caused by Sewer Pipe Cracks. Appl. Sci. 8, 679 (2018). https://doi.org/10.3390/app8050679
13. Mao, J., Wang, Y., Zhang, H., Jing, X.: Study on the Influence of Urban Water Supply Pipeline Leakage on the Scouring Failure Law of Cohesive Soil Subgrade. Water. 16, 93 (2023). https://doi.org/10.3390/w16010093
14. Dastpak, P., Sousa, R.L., Dias, D.: Soil Erosion Due to Defective Pipes: A Hidden Hazard Beneath Our Feet. Sustainability. 15, 8931 (2023). https://doi.org/10.3390/su15118931
15. Cui, J., Liu, F., Chen, R., Wang, S., Pu, C., Zhao, X.: Effects of Internal Pressure on Urban Water Supply Pipeline Leakage-Induced Soil Subsidence Mechanisms. Geofluids. 2024, 1–16 (2024). https://doi.org/10.1155/2024/9577375
16. Liu, J.-C., Wang, Z.-Y., Tan, Y., Cao, Y.-C.: Failure evolution and mechanism of ground collapse due to exfiltration of shallowly buried water pipeline. Eng. Fail. Anal. 162, 108390 (2024). https://doi.org/10.1016/j.engfailanal.2024.108390
17. Guo, J., Zhang, Y., Cheng, Y., Zhang, X., Shi, H., Zheng, J., Ma, Y.: Study on urban ground collapse induced by defective pipelines based on physical model experiments and numerical simulation. Sci. Rep. 15, 6085 (2025). https://doi.org/10.1038/s41598-025-90146-5
18. Chao, H., Tan, Y., Su, Z.-K.: Ground failure and soil erosion caused by bursting of buried water pipeline: experimental and numerical investigations. Eng. Fail. Anal. 167, 108965 (2025). https://doi.org/10.1016/j.engfailanal.2024.108965
19. Wang, Z.-Y., Liu, J.-C., Tan, Y., Long, Y.-Y.: Experimental and numerical investigation on internal erosion induced by infiltration of defective buried pipe. Bull. Eng. Geol. Environ. 84, 38 (2025). https://doi.org/10.1007/s10064-024-04073-2

20. Ali, H., Choi, J.: A Review of Underground Pipeline Leakage and Sinkhole Monitoring Methods Based on Wireless Sensor Networking. Sustainability. 11, 4007 (2019). https://doi.org/10.3390/su11154007
21. Tufano, R., Guerriero, L., Annibali Corona, M., Bausilio, G., Di Martire, D., Nisio, S., Calcaterra, D.: Anthropogenic sinkholes of the city of Naples, Italy: an update. Nat. Hazards. 112, 2577–2608 (2022). https://doi.org/10.1007/s11069-022-05279-x
22. ANSA: Voragine a Napoli, causa rottura condotta, https://www.ansa.it/campania/notizie/2015/02/22/voragine-a-napolicausa-rottura-condotta_f7321980-2a75-4825-a149-028ddce5d9a2.html, (2015)
23. Covelli, C., Cozzolino, L., Cimorelli, L., Della Morte, R., Pianese, D.: Optimal Location and Setting of PRVs in WDS for Leakage Minimization. Water Resour. Manag. 30, 1803–1817 (2016). https://doi.org/10.1007/s11269-016-1252-7
24. Casillas, M., Puig, V., Garza-Castañón, L., Rosich, A.: Optimal Sensor Placement for Leak Location in Water Distribution Networks Using Genetic Algorithms. Sensors. 13, 14984–15005 (2013). https://doi.org/10.3390/s131114984
25. Steffelbauer, D.B., Fuchs-Hanusch, D.: Efficient Sensor Placement for Leak Localization Considering Uncertainties. Water Resour. Manag. 30, 5517–5533 (2016). https://doi.org/10.1007/s11269-016-1504-6
26. Cugueró-Escofet, M.À., Puig, V., Quevedo, J.: Optimal pressure sensor placement and assessment for leak location using a relaxed isolation index: Application to the Barcelona water network. Control Eng. Pract. 63, 1–12 (2017). https://doi.org/10.1016/j.conengprac.2017.03.003
27. Li, J., Wang, C., Qian, Z., Lu, C.: Optimal sensor placement for leak localization in water distribution networks based on a novel semi-supervised strategy. J. Process Control. 82, 13–21 (2019). https://doi.org/10.1016/j.jprocont.2019.08.001
28. Hu, Z., Chen, W., Chen, B., Tan, D., Zhang, Y., Shen, D.: Robust Hierarchical Sensor Optimization Placement Method for Leak Detection in Water Distribution System. Water Resour. Manag. 35, 3995–4008 (2021). https://doi.org/10.1007/s11269-021-02922-3
29. Cheng, M., Li, J.: Optimal sensor placement for leak location in water distribution networks: A feature selection method combined with graph signal processing. Water Res. 242, 120313 (2023). https://doi.org/10.1016/j.watres.2023.120313
30. van Gemert, J.J.H., Breschi, V., Yntema, D.R., Keesman, K.J., Lazar, M.: Scalable Sensor Placement for Cyclic Networks with Observability Guarantees: Application to Water Distribution Networks, https://arxiv.org/abs/2508.13604, (2025)
31. Batzella, E., Ferrarese, G., Malavasi, S.: Sensor Placement for Rupture Detection Using a Continuous Monitoring Strategy. In: The 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024). p. 91. MDPI (2024)
32. Forconi, E., Kapelan, Z., Ferrante, M., Mahmoud, H., Capponi, C.: Risk-based sensor placement methods for burst/leak detection in water distribution systems. Water Supply. 17, 1663–1672 (2017). https://doi.org/10.2166/ws.2017.069
33. Hu, Z., Chen, W., Tan, D., Chen, B., Shen, D.: Multi-objective and risk-based optimal sensor placement for leak detection in a water distribution system. Environ. Technol. Innov. 28, 102565 (2022). https://doi.org/10.1016/j.eti.2022.102565
34. Parajuli, U., Magar, B.A., Ghimire, A.B., Shin, S.: Sensor Placement for the Classification of Multiple Failure Types in Urban Water Distribution Networks. Urban Sci. 9, 413 (2025). https://doi.org/10.3390/urbansci9100413
35. Medio, G., Varra, G., İnan, Ç.A., Cozzolino, L., Della Morte, R.: Sinkhole Risk-Based Sensor Placement for Leakage Localization in Water Distribution Networks with a Data-Driven Approach. Sustainability. 16, 5246 (2024). https://doi.org/10.3390/su16125246
36. Crichton, D.: The Risk Triangle. Tudor Rose, London (1999)

37. Bianchini, S., Confuorto, P., Intrieri, E., Sbarra, P., Di Martire, D., Calcaterra, D., Fanti, R.: Machine learning for sinkhole risk mapping in Guidonia-Bagni di Tivoli plain (Rome), Italy. Geocarto Int. 37, 16687–16715 (2022). https://doi.org/10.1080/10106049.2022.2113455
38. Zhang, Y., Jiao, Y.-Y., He, L.-L., Tan, F., Zhu, H.-M., Wei, H.-L., Zhang, Q.-B.: Susceptibility mapping and risk assessment of urban sinkholes based on grey system theory. Tunn. Undergr. Space Technol. 152, 105893 (2024). https://doi.org/10.1016/j.tust.2024.105893
39. Intrieri, E., Confuorto, P., Bianchini, S., Rivolta, C., Leva, D., Gregolon, S., Buchignani, V., Fanti, R.: Sinkhole risk mapping and early warning: the case of Camaiore (Italy). Front. Earth Sci. 11, 1172727 (2023). https://doi.org/10.3389/feart.2023.1172727
40. Barton, N.A., Farewell, T.S., Hallett, S.H., Acland, T.F.: Improving pipe failure predictions: Factors affecting pipe failure in drinking water networks. Water Res. 164, 114926 (2019). https://doi.org/10.1016/j.watres.2019.114926
41. Hussein Farh, H.M., Ben Seghier, M.E.A., Taiwo, R., Zayed, T.: Analysis and ranking of corrosion causes for water pipelines: a critical review. Npj Clean Water. 6, 65 (2023). https://doi.org/10.1038/s41545-023-00275-5
42. Muddassir, M., Zayed, T., Taiwo, R., Ben Seghier, M.E.A.: Advancing the analysis of water pipe failures: a probabilistic framework for identifying significant factors. Sci. Rep. 14, 19218 (2024). https://doi.org/10.1038/s41598-024-69855-w
43. Philip, B.E., Aljassmi, H.: The Relevance of Water Pipe Deterioration Prediction Models: A Review. Int. J. Sci. Technol. Res. 9, (2020)
44. Sinaei, A., Dziedzic, R., Creaco, E.: Holistic Assessment of Social, Environmental and Economic Impacts of Pipe Breaks: The Case Study of Vancouver. Water. 17, 252 (2025). https://doi.org/10.3390/w17020252
45. Wéber, R., Huzsvár, T., Hős, C.: Vulnerability analysis of water distribution networks to accidental pipe burst. Water Res. 184, 116178 (2020). https://doi.org/10.1016/j.watres.2020.116178
46. Wéber, R., Huzsvár, T., Hős, C.: Vulnerability of water distribution networks with real-life pipe failure statistics. Water Supply. 22, 2673–2682 (2022). https://doi.org/10.2166/ws.2021.447
47. Anglian Water: Drinking Water Network Failure Images., (2018)
48. U.S. Environmental Protection Agency (EPA): EPANET 2.2 User Manual. Office of Research and Development, U.S. Environmental Protection Agency, Washington, D.C. (2020)
49. Wagner, J.M., Shamir, U., Marks, D.H.: Water Distribution Reliability: Simulation Methods. J. Water Resour. Plan. Manag. 114, 276–294 (1988). https://doi.org/10.1061/(ASCE)0733-9496(1988)114:3(276)
50. Todini, E., Pilati, S.: Computer Applications in Water Supply: Volume 1 — Systems Analysis and Simulation. Gradient Algorithm Anal. Pipe Netw. (1988)
51. Kyriakou, M.S., Demetriades, M., Vrachimis, S.G., Eliades, D.G., Polycarpou, M.M.: EPyT: An EPANET-Python Toolkit for Smart Water NetworkSimulations. J. Open Source Softw. 8, 5947 (2023). https://doi.org/10.21105/joss.05947
52. Klise, K.A., Bynum, M., Moriarty, D., Murray, R.: A software framework for assessing the resilience of drinking water systems to disasters with an example earthquake case study. Environ. Model. Softw. 95, 420–431 (2017). https://doi.org/10.1016/j.envsoft.2017.06.022
53. Klise, K.A.; Murray, R.; Haxton, T.: An overview of the Water Network Tool for Resilience (WNTR). (2018)
54. Klise, K.A.; Hart, D.B.; Bynum, M.; Hogge, J.; Haxton, T.; Murray, R.; Burkhardt, J.: Water Network Tool for Resilience (WNTR) User Manual: Version 0.2.3. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC (2020)
55. Crowl, D.A.; Louvar, J.F.: Chemical Process Safety: Fundamentals with Applications. Prentice Hall, Upper Saddle River, NJ (2001)
56. Lambert, Allan: What do we know about pressure-leakage relationships in distribution systems. Presented at the Proceedings of the IWA International Specialised Conference: System Approach to Leakage Control and Water Distribution Systems Management May 16 (2001)

57. El-Zahab, S., Zayed, T.: Leak detection in water distribution networks: an introductory overview. Smart Water. 4, 5 (2019). https://doi.org/10.1186/s40713-019-0017-x
58. Romero-Ben, L., Alves, D., Blesa, J., Cembrano, G., Puig, V., Duviella, E.: Leak detection and localization in water distribution networks: Review and perspective. Annu. Rev. Control. 55, 392–419 (2023). https://doi.org/10.1016/j.arcontrol.2023.03.012
59. Pudar, R.S., Liggett, J.A.: Leaks in Pipe Networks. J. Hydraul. Eng. 118, 1031–1046 (1992). https://doi.org/10.1061/(ASCE)0733-9429(1992)118:7(1031)
60. Casillas Ponce, M.V., Garza Castañón, L.E., Cayuela, V.P.: Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities. J. Hydroinformatics. 16, 649–670 (2014). https://doi.org/10.2166/hydro.2013.019
61. Ramon Perez; Gerard Sanz; Vicenc Puig; Joseba Quevedo; Miquel Angel Cuguero Escofet; Fatiha Nejjari: Leak Localization in Water Networks: A Model-Based Methodology Using Pressure Sensors Applied to a Real Network in Barcelona [Applications of Control]. IEEE Control Syst. 34, 24–36 (2014). https://doi.org/10.1109/MCS.2014.2320336
62. Zhang, H., Wang, L.: Leak detection in water distribution systems using Bayesian theory and Fisher's law. Trans. Tianjin Univ. 17, 181–186 (2011). https://doi.org/10.1007/s12209-011-1594-4
63. Sanz, G., Perez, R., Escobet, A.: Leakage localization in water networks using fuzzy logic. In: 2012 20th Mediterranean Conference on Control & Automation (MED). pp. 646–651. IEEE, Barcelona, Spain (2012)
64. Bicik, J., Kapelan, Z., Makropoulos, C., Savić, D.A.: Pipe burst diagnostics using evidence theory. J. Hydroinformatics. 13, 596–608 (2011). https://doi.org/10.2166/hydro.2010.201
65. Sun, C., Parellada, B., Puig, V., Cembrano, G.: Leak Localization in Water Distribution Networks Using Pressure and Data-Driven Classifier Approach. Water. 12, 54 (2019). https://doi.org/10.3390/w12010054
66. Ares-Milián, M.J., Quiñones-Grueiro, M., Verde, C., Llanes-Santiago, O.: A Leak Zone Location Approach in Water Distribution Networks Combining Data-Driven and Model-Based Methods. Water. 13, 2924 (2021). https://doi.org/10.3390/w13202924
67. Wachla, D., Przystalka, P., Moczulski, W.: A Method of Leakage Location in Water Distribution Networks using Artificial Neuro-Fuzzy System. IFAC-Pap. 48, 1216–1223 (2015). https://doi.org/10.1016/j.ifacol.2015.09.692
68. Li, J., Zheng, W., Lu, C.: An Accurate Leakage Localization Method for Water Supply Network Based on Deep Learning Network. Water Resour. Manag. 36, 2309–2325 (2022). https://doi.org/10.1007/s11269-022-03144-x
69. Romano, M., Woodward, K., Kapelan, Z.: Statistical Process Control Based System for Approximate Location of Pipe Bursts and Leaks in Water Distribution Systems. Procedia Eng. 186, 236–243 (2017). https://doi.org/10.1016/j.proeng.2017.03.235
70. Laucelli, D., Romano, M., Savić, D., Giustolisi, O.: Detecting anomalies in water distribution networks using EPR modelling paradigm. J. Hydroinformatics. 18, 409–427 (2016). https://doi.org/10.2166/hydro.2015.113
71. Alves, D., Blesa, J., Duviella, E., Rajaoarisoa, L.: Robust Data-Driven Leak Localization in Water Distribution Networks Using Pressure Measurements and Topological Information. Sensors. 21, 7551 (2021). https://doi.org/10.3390/s21227551
72. Hutton, C., Kapelan, Z.: Real-time Burst Detection in Water Distribution Systems Using a Bayesian Demand Forecasting Methodology. Procedia Eng. 119, 13–18 (2015). https://doi.org/10.1016/j.proeng.2015.08.847
73. Badillo, S., Banfai, B., Birzele, F., Davydov, I.I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., Zhang, J.D.: An Introduction to Machine Learning. Clin. Pharmacol. Ther. 107, 871–885 (2020). https://doi.org/10.1002/cpt.1796
74. Breiman, L.: Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat. Sci. 16, (2001). https://doi.org/10.1214/ss/1009213726
75. Spector, A.: Data Science and AI in Context: Summary and Insights. Harv. Data Sci. Rev. 6, (2024). https://doi.org/10.1162/99608f92.cdebd845

76. Sarker, I.H.: Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput. Sci. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
77. Song, Y.-Y., Lu, Y.: Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry. 27, 130–135 (2015). https://doi.org/10.11919/j.issn.1002-0829.215044
78. Blockeel, H., Devos, L., Frénay, B., Nanfack, G., Nijssen, S.: Decision trees: from efficient prediction to responsible AI. Front. Artif. Intell. 6, 1124553 (2023). https://doi.org/10.3389/frai.2023.1124553
79. Pedregosa, Fabian; Varoquaux, Gaël; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier; Blondel, Mathieu; Prettenhofer, Peter; Weiss, Ron; Dubourg, Vincent; Vanderplas, Jake; Passos, Alexandre; Cournapeau, David; Brucher, Matthieu; Perrot, Matthieu; Duchesnay, Édouard: Scikit-learn: Machine Learning in Python. 12, 2825–2830 (2011)
80. Hu, Z., Chen, B., Chen, W., Tan, D., Shen, D.: Review of model-based and data-driven approaches for leak detection and location in water distribution systems. Water Supply. 21, 3282–3306 (2021). https://doi.org/10.2166/ws.2021.101
81. Pudar, R.S., Liggett, J.A.: Leaks in Pipe Networks. J. Hydraul. Eng. 118, 1031–1046 (1992). https://doi.org/10.1061/(ASCE)0733-9429(1992)118:7(1031)
82. Pérez, R., Puig, V., Pascual, J., Peralta, A., Landeros, E., Jordanas, Ll.: Pressure sensor distribution for leak detection in Barcelona water distribution network. Water Supply. 9, 715–721 (2009). https://doi.org/10.2166/ws.2009.372
83. Pérez, R., Puig, V., Pascual, J., Quevedo, J., Landeros, E., Peralta, A.: Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks. Control Eng. Pract. 19, 1157–1167 (2011). https://doi.org/10.1016/j.conengprac.2011.06.004
84. Pérez, R., Cugueró, M.-A., Cugueró, J., Sanz, G.: Accuracy Assessment of Leak Localisation Method Depending on Available Measurements. Procedia Eng. 70, 1304–1313 (2014). https://doi.org/10.1016/j.proeng.2014.02.144
85. Leak Localization in Water Networks: A Model-Based Methodology Using Pressure Sensors Applied to a Real Network in Barcelona [Applications of Control]. IEEE Control Syst. 34, 24–36 (2014). https://doi.org/10.1109/MCS.2014.2320336
86. Pérez, R., Cugueró, J., Blesa, J., Cugueró, M.A., Sanz, G.: Uncertainty effect on leak localisation in a DMA. In: 2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol). pp. 319–324. IEEE, Barcelona, Spain (2016)
87. Blesa, J., Pérez, R.: Modelling uncertainty for leak localization in Water Networks. IFAC-Pap. 51, 730–735 (2018). https://doi.org/10.1016/j.ifacol.2018.09.656
88. Bartkowska, I., Wysocki, Ł., Zajkowski, A., Tuz, P.: Comparative Analysis of Leak Detection Methods Using Hydraulic Modelling and Sensitivity Analysis in Rural and Urban–Rural Areas. Sustainability. 16, 7405 (2024). https://doi.org/10.3390/su16177405
89. Blank, J., Deb, K.: Pymoo: Multi-Objective Optimization in Python. IEEE Access. 8, 89497–89509 (2020). https://doi.org/10.1109/ACCESS.2020.2990567
90. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. The MIT Press (1992)
91. Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. Multimed. Tools Appl. 80, 8091–8126 (2021). https://doi.org/10.1007/s11042-020-10139-6
92. Sastry, K., Goldberg, D., Kendall, G.: Genetic Algorithms. In: Burke, E.K. and Kendall, G. (eds.) Search Methodologies. pp. 97–125. Springer US, Boston, MA (2005)
93. Alam, T., Qamar, S., Dixit, A., Benaida, M.: Genetic Algorithm: Reviews, Implementations, and Applications. (2020). https://doi.org/10.48550/ARXIV.2007.12673
94. Londe, M.A., Pessoa, L.S., Andrade, C.E., Resende, M.G.C.: Biased random-key genetic algorithms: A review. Eur. J. Oper. Res. 321, 1–22 (2025). https://doi.org/10.1016/j.ejor.2024.03.030
95. Vie, A., Kleinnijenhuis, A.M., Farmer, D.J.: Qualities, challenges and future of genetic algorithms: a literature review, https://arxiv.org/abs/2011.05277, (2020)
96. Drachal, K., Pawłowski, M.: A Review of the Applications of Genetic Algorithms to Forecasting Prices of Commodities. Economies. 9, 6 (2021). https://doi.org/10.3390/economies9010006

97. Manning, T., Sleator, R.D., Walsh, P.: Naturally selecting solutions: the use of genetic algorithms in bioinformatics. Bioengineered. 4, 266–278 (2013). https://doi.org/10.4161/bioe.23041
98. Maharana, K., Mondal, S., Nemade, B.: A review: Data pre-processing and data augmentation techniques. Glob. Transit. Proc. 3, 91–99 (2022). https://doi.org/10.1016/j.gltp.2022.04.020
99. Kurita, T.: Principal Component Analysis (PCA). In: Computer Vision. pp. 1–4. Springer International Publishing, Cham (2020)
100. Istituto Nazionale di Statistica (ISTAT): Basi territoriali e variabili censuarie, https://www.istat.it/notizia/basi-territoriali-e-variabili-censuarie/, (2025)
101. Città Metropolitana di Napoli: Sistema Informativo Territoriale (SIT) – Città Metropolitana di Napoli, https://sit.cittametropolitana.na.it/index.php
102. Google LLC: Google Earth Pro, https://earth.google.com/web/
103. Google LLC: Google Maps, https://www.google.com/maps
104. Ministero delle Infrastrutture e dei Trasporti: Norme Tecniche per le Costruzioni (NTC 2018), https://www.gazzettaufficiale.it/eli/gu/2018/02/20/42/so/8/sg/pdf/, (2018)
105. Quiñones-Grueiro, M., Bernal-de Lázaro, J.M., Verde, C., Prieto-Moreno, A., Llanes-Santiago, O.: Comparison of Classifiers for Leak Location in Water Distribution Networks. IFAC-Pap. 51, 407–413 (2018). https://doi.org/10.1016/j.ifacol.2018.09.609
106. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. (2013). https://doi.org/10.48550/ARXIV.1309.0238
107. Vrachimis, S.G., Eliades, D.G., Taormina, R., Kapelan, Z., Ostfeld, A., Liu, S., Kyriakou, M., Pavlou, P., Qiu, M., Polycarpou, M.M.: Battle of the Leakage Detection and Isolation Methods. J. Water Resour. Plan. Manag. 148, 04022068 (2022). https://doi.org/10.1061/(ASCE)WR.1943-5452.0001601
108. Vrachimis, S.G., Timotheou, S., Eliades, D.G., Polycarpou, M.M.: Iterative Hydraulic Interval State Estimation for Water Distribution Networks. J. Water Resour. Plan. Manag. 145, 04018087 (2019). https://doi.org/10.1061/(ASCE)WR.1943-5452.0001011
109. Thomas Brinkhoff: CityPopulation.de, https://www.citypopulation.de/
110. Statistical Service of Cyprus: Statistical Service of Cyprus (CyStat), https://www.cystat.gov.cy/en/
111. European Environment Agency (EEA): Urban Atlas – Copernicus Land Monitoring Service, https://land.copernicus.eu/en/products/urban-atlas?tab=technical_summary
112. European Environment Agency: Urban Atlas Land Cover/Land Use 2018 (vector), Europe, 6-yearly, Jul. 2021, https://sdi.eea.europa.eu/catalogue/copernicus/api/records/fb4dffa1-6ceb-4cc0-8372-1ed354c285e6?language=all, (2020)
113. Gould, S.J.F., Davis, P., Beale, D.J., Marlow, D.R.: Failure analysis of a PVC sewer pipeline by fractography and materials characterization. Eng. Fail. Anal. 34, 41–50 (2013). https://doi.org/10.1016/j.engfailanal.2013.07.009
114. Bruaset, S., Sægrov, S.: An Analysis of the Potential Impact of Climate Change on the Structural Reliability of Drinking Water Pipes in Cold Climate Regions. Water. 10, 411 (2018). https://doi.org/10.3390/w10040411
115. Martínez-Codina, Á., Castillo, M., González-Zeas, D., Garrote, L.: Pressure as a predictor of occurrence of pipe breaks in water distribution networks. Urban Water J. 13, 676–686 (2016). https://doi.org/10.1080/1573062X.2015.1024687
116. Moslehi, I., Jalili_Ghazizadeh, M.: Pressure-Pipe Breaks Relationship in Water Distribution Networks: A Statistical Analysis. Water Resour. Manag. 34, 2851–2868 (2020). https://doi.org/10.1007/s11269-020-02587-4
117. Konstantinou, C., Jara-Arriagada, C., Stoianov, I.: Investigating the Impact of Cumulative Pressure-Induced Stress on Machine Learning Models for Pipe Breaks. Water Resour. Manag. 38, 603–619 (2024). https://doi.org/10.1007/s11269-023-03687-7
118. Gruppo CAP: Regolamento del Servizio Idrico Integrato. (2022)
119. SMAT (Società Metropolitana Acque Torino): Il regolamento del servizio idrico integrato. (2018)

120. Vrachimis, S.G., Eliades, D.G., Taormina, R., Kapelan, Z., Ostfeld, A., Liu, S., Kyriakou, M., Pavlou, P., Qiu, M., Polycarpou, M.M.: Battle of the Leakage Detection and Isolation Methods. J. Water Resour. Plan. Manag. 148, 04022068 (2022). https://doi.org/10.1061/(ASCE)WR.1943-5452.0001601
121. Rivera, M.M., Ochoa-Zezzatti, A., Serna, S.P.: Embedded system for model characterization developing intelligent controllers in industry 4.0. In: Artificial Intelligence and Industry 4.0. pp. 57–91. Elsevier (2022)