Real-Time Leak Detection and Localization in Water Distribution Networks via Mahalanobis Space Transformation

Oleg Melnikov[1], Yurii Dorofieiev[1], Yurii Shakhnovskiy[1], Huy Truong[2], and Victoria Degeler[3]

[1] Institute of Computer Science and Information Technologies, National Technical University "Kharkiv Polytechnic Institute", Ukraine
Oleg.Melnikov@khpi.edu.ua,
[2] Bernoulli Institute, University of Groningen, The Netherlands
[3] Informatics Institute, University of Amsterdam, The Netherlands,
v.o.degeler@uva.nl

Abstract. This paper presents a statistical framework for leak detection and localization in water distribution networks based on real-time SCADA measurements. The proposed approach relies on a Mahalanobis transformation of sensor data, which explicitly accounts for spatial correlations between measurements and provides a whitening of the observation space. Under normal operating conditions, the resulting quadratic form follows Hotelling's T^2 distribution, enabling systematic detection of both abrupt and incipient leaks with explicit control of the false alarm rate. The model can be trained either on a hydraulic model or from historical SCADA data representing normal operation. Leak localization is performed by matching transformed sensor observations with precomputed hydraulic simulations of pipe-specific leak scenarios. Simulation studies demonstrate improved detection responsiveness and localization accuracy compared to alternative approaches.

Keywords: anomaly detection, leak localization, water distribution network, Mahalanobis distance, Hotelling statistic

1 Introduction

Losses of drinking water caused by leaks and pipe bursts in water distribution networks (WDNs) represent a serious challenge for water utilities, leading to reduced revenues, infrastructure degradation, and adverse environmental impacts. Given rapidly increasing global water demand and energy costs, the problem of leak detection and localization in WDNs has attracted considerable research attention [7,8]. The traditional approach to addressing this problem involves two stages: (i) leak detection, i.e. identifying the presence of a leak through the observation of significant deviations from normal behavior in network operations; and (ii) leak localization, i.e. determining the approximate location of the leak. Leaks result in increased water flow and reduced pipe pressure. As most pipes are underground, operators of WDNs primarily rely on flow and pressure sensors to assess the system state. However, leak detection and localization are substantially complicated by multiple sources of uncertainty in WDN operations, such as demand fluctuations, pipe condition, and sensor measurement errors [6].

Numerous techniques for leak detection and localization in WDNs have been introduced, particularly in the context of the "Battle of the Leakage Detection and Isolation Methods" (BattLeDIM) competition [10]. They range from on-site inspections using dedicated hardware to software-based processing of data collected from IoT sensors. Software methods are commonly classified into three categories: (i) model-based methods that explicitly rely on a hydraulic model of the network; (ii) data-driven methods that do not require a network model; and (iii) hybrid approaches that combine model-based and data-driven techniques.

The core principle of model-based leak detection consists of comparing measurements obtained from Supervisory Control And Data Acquisition (SCADA) systems with reference values simulated under nominal operating conditions by hydraulic models and corresponding software such as EPANET and WNTR. Leak localization is performed by comparing SCADA data with simulated leak scenarios at different locations in the network and minimizing the discrepancy between simulated and measured data according to suitable similarity metrics. The main limitation of this approach is that the hydraulic model must be calibrated so it can accurately reproduce the system's observed behavior.

Data-driven methods focus on the knowledge mined from historical SCADA records. They employ machine learning models that are trained on past data that characterizes normal system behavior [7]. Leaks are then identified by examining discrepancies between measured and model-predicted system variables. The effectiveness of these methods strongly relies on the availability of large, high-quality training datasets. While data representing normal operating conditions are relatively easy to collect, obtaining extensive datasets on leak events at various locations is far more challenging. Consequently, many works employ hybrid strategies, combining data-driven techniques for leak detection with model-based methods for leak localization [1,9].

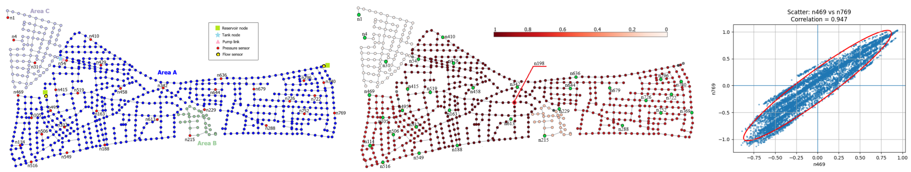
Our study departs from the existing research in several important aspects. We apply a Mahalanobis transformation to the sensor measurements, which removes spatial correlations and puts them on a common scale. Under mild regularity conditions, the quadratic form of these transformed measurements follows a known (Hotelling) distribution. This property enables us to pose leak detection as a standard hypothesis testing problem, providing explicit control over the false alarm rate and ensuring that the results are statistically interpretable.

2 Materials and Methods

Since the selection of the proposed methods was influenced by the characteristics of the available data, a brief description of the analyzed dataset is provided first.

The analysis employed the L-Town benchmark model from BattLeDIM [10]. It describes the WDN of a small city partitioned into three district metered areas

(DMAs). The network is monitored by 33 pressure sensors, three flow meters, and a water level sensor installed in the storage tank (Fig. 1a). Numerous leak scenarios were simulated using a hydraulic model, and the corresponding SCADA data representing two years of WDN operation were subsequently collected.



(a) (b) (c)

Fig. 1: L-Town features: (a) – DMA partitioning and sensor locations; (b) – correlation of pressure values with the centrally located node `n198`; (c) – scatter plot of pressure measurements from two spatially distant sensors `n469` and `n769` in Area A based on 2019 SCADA data.

This dataset features strong spatial correlation between measurements among sensors located within the same DMA, while the hydraulic behavior of different DMAs is practically independent. This effect is evident both in hydraulic simulations (Fig. 1b) and in time series of SCADA data (Fig. 1c). Such correlations substantially complicate leak detection and localization, as they reduce the effective dimensionality of the data and violate the independence assumptions underlying many classical detection methods.

One approach that both makes the sensor data comparable on a common scale and explicitly incorporates their correlation structure is to use the Mahalanobis distance. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^s$ are two vectors associated with a probability distribution characterized by the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the Mahalanobis distance between them is defined as:

$$D_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}. \tag{1}$$

If $\boldsymbol{\Sigma}$ is the identity matrix, then the Mahalanobis distance coincides with the Euclidean distance $D_E(\mathbf{x}, \mathbf{y})$. If $\boldsymbol{\Sigma}$ is diagonal, formula (1) reduces to the normalized Euclidean distance:

$$D_{NE}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{s} \left(\frac{x_i - y_i}{\sigma_i}\right)^2}, \tag{2}$$

where σ_i is the standard deviation of the i-th component.

Alternatively, one can apply a Mahalanobis (or whitening) transformation

$$\mathbf{z} = M(\mathbf{x}) = \mathbf{W}(\mathbf{x} - \boldsymbol{\mu}), \text{ with } \mathbf{W}^\top \mathbf{W} = \boldsymbol{\Sigma}^{-1}, \tag{3}$$

which yields a whitened vector $\mathbf{z}$ having zero mean and an identity covariance matrix. Using this notation, the Mahalanobis distance can be interpreted as a standard Euclidean distance between the whitened images of $\mathbf{x}$ and $\mathbf{y}$.

From an anomaly detection standpoint, the distance $D_M(\mathbf{x}, \boldsymbol{\mu})$ is particularly important, as it quantifies how far the observation $\mathbf{x}$ lies from the center of the distribution. If $\mathbf{x}$ is a multivariate normal vector, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the squared Mahalanobis distance $Z^2 = D_M^2(\mathbf{x}, \boldsymbol{\mu})$ follows a chi-squared distribution with s degrees of freedom. Thus, if we interpret $\mathbf{x}$ as a random vector of sensor measurements under normal operational mode, then improbably high values of Z^2 may indicate the emergence of a leak or other abnormal condition[1].

In practice, true values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and should be replaced with their estimates $\bar{\mathbf{x}}, \mathbf{S}$, obtained from a sample $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \in \mathbb{R}^s$:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum\nolimits_{i=1}^{n} \mathbf{x}^{(i)}, \qquad \mathbf{S} = \frac{1}{n-1} \sum\nolimits_{i=1}^{n} (\mathbf{x}^{(i)} - \bar{\mathbf{x}})(\mathbf{x}^{(i)} - \bar{\mathbf{x}})^{\mathsf{T}}. \tag{4}$$

The squared Mahalanobis distance computed with these estimates

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) = M(\mathbf{x})^{\mathsf{T}} M(\mathbf{x}) \tag{5}$$

follows a Hotelling's T^2 distribution with (n, s) degrees of freedom [2].

These properties offer a robust statistical foundation for leak detection with explicit control over the false alarm rate, while the whitening of sensor data also facilitates leak localization.

3 Implementation and Validation

Model implementation requires estimating the moments of the sensor measurement vector under normal operating conditions. Since sensor measurements are affected by diurnal and seasonal demand variations, the data is partitioned into corresponding temporal clusters. The observations can also be organized into spatial clusters when sensors in those regions behave independently from the rest. For L-Town the measurements were divided into spatio-temporal groups (STGs) determined by a combination of DMA and the time-of-day cluster. For each sub-group g, the distribution moments (mean vectors and covariance matrices) are estimated separately. The estimation procedure depends on the underlying data source. When relying on a hydraulic WDN model, the mean vectors $\boldsymbol{\mu}_g$ are set to the nominal values obtained from simulation software. The covariance matrices for each STG are estimated using Eq. (4) by averaging over the number of observations in this group and the total number of simulation runs with randomized parameters which reflect the historical variability. Whitening matrices $\mathbf{W}_g$ are then computed using the Cholesky decomposition of the corresponding covariance matrix. Time series data offers only one historical realization, limiting temporal clustering granularity and risking contamination by undetected anomalies [5]. For formal statistical validity, the data distribution within each STG should be checked for normality. However, for large sensor counts, the Hotelling's T^2 statistic remains effective even if some normality tests fail, as the aggregation in (5) mitigates moderate departures from Gaussian assumptions.

[1]Other abnormal conditions may arise from sensor malfunctions or unobserved demand (e.g., unmetered water withdrawals). The methodology presented in [5] discusses approaches for distinguishing these anomalies from actual leak events.

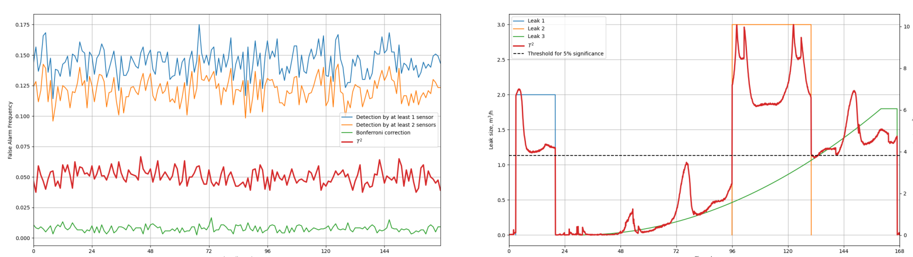
3.1 Leak Detection

After the model is trained, leak detection is straightforward. For each SCADA observation, an appropriate STG group g with its associated whitening matrix $\mathbf{W}_g$ is selected. The sensor data are whitened using the formula (3) and the Hotelling's T^2 statistic (5) is computed. The obtained value is compared with a critical threshold $\theta_g(\alpha) = H_{1-\alpha}(s, n_g - s)$, where α is a target significance level, n_g is the number of observations used for moment estimation in g-th STG, and $H_q(d_1, d_2)$ is the q^{th} quantile of the Hotelling distribution with (d_1, d_2) degrees of freedom. Condition $T^2 > \theta_g(\alpha)$ then signals the presence of an anomaly.

To validate the model, the L-Town WDN was simulated over one week using WNTR/EPANET. To represent uncertainty, a multiplicative Gaussian noise with a standard deviation of 10% was superimposed on the nodal demands.

First, we study the false alarms during normal operations. Tuning for a specific false alarm rate is challenging because system-level error rates differ substantially from individual sensor rates. The leak detection results of BattLeDIM [10] exhibit a high rate of false alarms. The traditional solution to this problem is the Bonferroni correction, which adjusts the significance threshold according to the number of sensors [2]. However, in the presence of strong correlations among sensor measurements, the Bonferroni correction is typically overly conservative, the nominal level, which can significantly reduce anomaly detection sensitivity.

Fig. 2a presents the false alarm rates obtained from 100 simulations for 4 leak-detection criteria: at least (i) one or (ii) two sensors exhibiting anomalous readings (iii) Bonferroni correction, and (iv) Hotelling's T^2 statistic. The rates of methods (i)-(iii) substantially differ from the intended 5% significance level, whereas the T^2-based detection method (iv) achieves a rate close to it.



(a) (b)

Fig. 2: Leak detection using T^2 statistics: (a) – false alarm rate for various detection methods; (b) – T^2 time series for operations with three simulated leaks

Fig. 2b illustrates the T^2 statistic with three simulated leaks of different magnitudes, one of which is time-varying. The dashed line corresponds to a 5% significance level. Two abrupt leaks are promptly identified, whereas the incipient leak becomes noticeable only after exceeding a certain threshold. The spikes are due to demand heterogeneity: leaks are considerably easier to detect at night when water consumption variance is much lower than during daytime.

The dynamics of the T^2 statistic can be exploited to detect emerging leaks even in the presence of ongoing leaks, an aspect discussed in more detail in [5].

3.2 Leak Localization

Leak localization can be formulated as an optimization problem. Let $\mathbf{x}_{\text{obs}} \in \mathbb{R}^s$ denote the observed sensor measurements vector, and $\mathbf{x}_j(v) \in \mathbb{R}^s$ represent the simulated sensor response matching a leak of magnitude v occurring in pipe j, obtained from hydraulic simulations. The localization task is expressed as

$$(\hat{j}, \hat{v}) = \arg\min_{j,\, v} D(\mathbf{x}_{\text{obs}}, \mathbf{x}_j(v))\,, \tag{6}$$

where $D(\mathbf{x}, \mathbf{y})$ denotes a dissimilarity measure defined in the sensor measurement space. Some choices of the distance metric D that have been proposed include:

- the standard Euclidean distance $D_E(\mathbf{x}, \mathbf{y})$ [3];
- a normalized Euclidean distance $D_{NE}(\mathbf{x}, \mathbf{y})$ given by (2) [9];
- a cosine distance $D_{\cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^\intercal \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$ [11,12].

Another approach involves extending the state space by constructing pairwise pressure sensor differentials $\Delta = (x_i - x_j)_{1 \le i < j \le s} \in \mathbb{R}^{s(s-1)/2}$, which are then compared using the Euclidean distance [4]. We refer to this method as Δ_E.

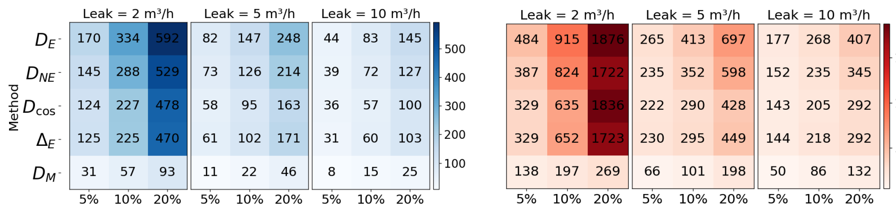
These metrics neglect the spatial correlations intrinsic to WDN sensor measurements within the same DMA. In contrast, the Mahalanobis distance (1) explicitly incorporates the inter-sensor correlation structure, thereby improving discrimination between competing leak scenarios.

To verify these considerations, two groups of experiments were conducted.

The objective of the first group of experiments was to assess the comparative performance of leak localization methods as a function of the leak magnitude and the noise level in the system. A leak was simulated sequentially in each of the 901 inter-junction pipes of the WDN, and an attempt was made to identify the pipe in which the leak occurred. The leak magnitude was assumed to be known a priori and was varied across three levels: 2, 5, and 10 m^3/h. As in the previous experiments, noise was modeled as multiplicative Gaussian noise applied to nodal demands, with intensity levels of 5%, 10%, and 20%.

Localization accuracy was evaluated using the geometric (Euclidean) distance e_g between the midpoints of the true pipe and the pipe identified by the localization method. Fig. 3 reports the mean value and the 95-th percentile of e_g distribution for various scenarios as heatmaps; darker shades indicate larger localization errors. The results are largely intuitive: for all considered methods, the localization error decreases as the leak magnitude increases and grows with increasing noise levels. For all scenarios, localization based on the Mahalanobis distance exhibits significantly higher accuracy compared to the alternative methods, although this advantage diminishes when the leak-to-noise ratio is high.

The second set of experiments focuses on leak localization using D_M metrics when the leak magnitude is not known in advance. Two search strategies were employed: (i) an exhaustive search over a coarse grid of candidate



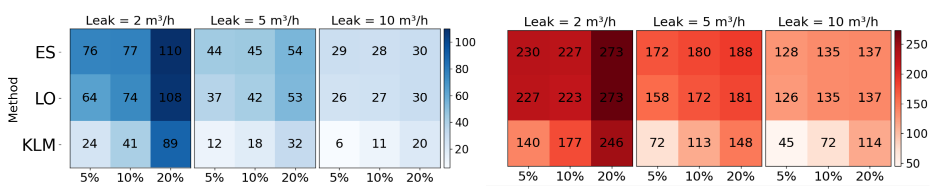
(a) (b)

Fig. 3: Geometric error of different localization methods, meters: (a) – mean value; (b) – 95%-th quantile

leak magnitudes (ES). Although computationally demanding, the grid nodes (pipe–magnitude pairs) can be precomputed and stored, ensuring efficient real-time execution; (ii) a grid-based search followed by a local optimization around the identified candidate (LO). The latter alternates between one-dimensional refinement of the leak magnitude and discrete updates of the pipe index within a k-neighborhood (k=5), using a coordinate descent scheme. The process continues until no further decrease of the objective function is observed.

As in previous experiments, leaks of different magnitudes were simulated in all pipes at various noise levels. The accuracy of leak localization is reported in Fig. 4. For reference, the localization errors obtained with known leak magnitudes are also reported (KLM). The mean leak magnitude estimation error was below 0.5 m^3/h.

Constructing the grid with a 1 m^3/h resolution required 1218 s on a machine equipped with an Intel i3-10100 3.6 GHz CPU and 8 GB of RAM. Subsequently, leak localization took on average 0.1 s with the ES method and 11 s with LO.



(a) (b)

Fig. 4: Geometric error of Mahalanobis-based localization with the unknown leak magnitude, meters: (a) – mean value; (b) – 95%-th quantile

The results indicate that the localization accuracy is only slightly degraded compared to the case where the leak magnitude is known. This confirms the practical applicability of the proposed approach.

4 Conclusion

This paper proposes a statistical framework for leak detection and localization in water distribution networks based on a Mahalanobis transformation of SCADA measurements. Simulation results demonstrate the advantages of the proposed approach over alternative methods in terms of responsiveness and precision. However, further validation using real-world data is required.

Future research will focus on practical deployment, including robustness to modeling uncertainties, incorporation of temporal dynamics, and adaptation to evolving operational conditions.

Acknowledgments This work has received support from the Dutch Research Council (NWO) and the National Research Foundation of Ukraine (NRFU), in the scope of the Digital Twin for Evolutionary Changes in Water Networks (DiTEC) project, file numbers 19454 and NWOPI.UKR.2023.001.

References

1. Ahn, J., Jung, D.: Hybrid statistical process control method for water distribution pipe burst detection. J. Water Resour. Plann. Manage. **145**(9), 198483109 (2019)
2. Bersimis, S., Psarakis, S., Panaretos, J.: Multivariate statistical process control charts: an overview. Qual. Reliab. Eng. Int. **23**(5), 517—-543 (2007)
3. Marzola, I., Alvisi, S., Franchini, M.: A comparison of model-based methods for leakage localization in water distribution systems. Water Resources Management **36**(14), 5711–5727 (2022). DOI 10.1007/s11269-022-03329-4
4. Mazzoni, F., Marsili, V., Alvisi, S., Franchini, M.: Detection and pre-localization of anomalous consumption events in water distribution networks through automated, pressure-based methodology. Water Resour. Indust. **31**, 100255 (2024)
5. Melnikov, O., Dorofieiev, Y., Shakhnovskiy, Y., Truong, H., Degeler, V.: A multivariate statistical framework for detection, classification and pre-localization of anomalies in water distribution networks. Expert Systems with Applications **313**, 131450 (2026). DOI 10.1016/j.eswa.2026.131450
6. Mohan Doss, P., Rokstad, M.M., Steffelbauer, D., Tscheikner-Gratl, F.: Uncertainties in different leak localization methods for water distribution networks: a review. Urban Water Journal **20**(8), 953–967 (2023). DOI 10.1080/1573062X.2023.2229301
7. Nimri, W., Wang, Y., Zhang, Z., Deng, C., Sellstrom, K.: Data-driven approaches and model-based methods for detecting and locating leaks in water distribution systems: a literature review. Neural Comput & Applic **35**, 11611–11623 (2023)
8. Romero-Ben, L., Alves, D., Blesa, J., Cembrano, G., Puig, V.: Leak localization in water distribution networks using data-driven and model-based approaches. J. Water Resour. Plann. Manage. **148**(5), 04022016 (2022)
9. Steffelbauer, D.B., Deuerlein, J., Gilbert, D., Abraham, E., Piller, O.: Pressure-leak duality for leak detection and localization in water distribution systems. J. Water Resour. Plann. Manage. **148**(3), 04021106 (2022)
10. Vrachimis, S.G., Eliades, D.G., Taormina, R., Kapelan, Z., Ostfeld, A., Liu, S., Kyriakou, M., ..., Polycarpou, M.M.: Battle of the leakage detection and isolation methods. Journal of Water Res. Plan. & Manage., **148**(12), 04022068 (2022)
11. Wan, X., Farmani, R., Keedwell, E.: Gradual leak detection in water distribution networks based on multistep forecasting strategy. J. Water Res. Plann. Manage. **149**(8), 04023035 (2023). DOI 10.1061/JWRMD5.WRENG-6001
12. Yi, K., Liu, X., Hu, W., Li, Y.: Pipeline anomaly detection based on KPCA and cosine distance prototypical network. In: IEEE International Conference on Industrial Technology (ICIT), pp. 1–6 (2025). DOI 10.1109/ICIT63637.2025.10965214