

Anomaly Detection at the DMA-Level via Isolation Forest [†]

Chiara Cincotta ^{1,*}, Lorenzo Pedroni ², Michele Lombardi ², Giuditta Nicoli ³ and Cristiana Bragalli ¹

¹ Department of Civil, Chemical, Environmental and Materials Engineering—DICAM, University of Bologna, 40136 Bologna, Italy; cristiana.bragalli@unibo.it

² Department of Computer Science and Engineering—DISI, University of Bologna, 40136 Bologna, Italy; lorenzo.pedroni3@studio.unibo.it (L.P.); michele.lombardi2@unibo.it (M.L.)

³ Hera S.p.A., 41122 Modena, Italy; giuditta.nicoli@gruppohera.it

* Correspondence: chiara.cincotta@unibo.it

[†] Presented at II International Conference on Challenges and Perspectives in Urban Water Management Systems (CSDU-CSSI DAYS 25), Trieste, Italy, 18–19 November 2025.

Abstract

This study applies the Isolation Forest (IF) algorithm to detect anomalies in a district metered area (DMA) in Emilia-Romagna, Italy. Multiple datasets are analyzed, including 15-min inflows, daily minima, and inflows excluding the consumption of a high-demand industrial user. Anomalies are cross-referenced with repair records to assess correlation with leaks and failures and a metric is defined to evaluate the algorithm performance across datasets. Results show that sensor malfunctions and communication anomalies can be effectively detected through the application of the IF algorithm. Regarding the detection of burst and leakage events, the automated analysis of daily minima is the most effective and removing industrial consumption enhances detection accuracy.

Keywords: water distribution systems; district metered area; anomaly detection; isolation forest

1. Introduction

In the past decade, the growing adoption of smart sensors alongside SCADA (Supervisory Control and Data Acquisition) systems has driven the development of advanced data analytics to improve the efficiency of drinking water services [1]. This advanced metering infrastructure generates large volumes of data, whose effective use depends on the timely detection of anomalies, arising from sensor malfunctions, communication failures, pipe breaks, or unauthorized water use.

The datasets used in this study are provided by Hera S.p.A, one of the main utility companies responsible for water supply and distribution in Emilia-Romagna, IT. For water distribution management, the network is divided into DMAs, where flow is measured at the inlet and outlet points at regular time intervals. This study considers a specific DMA (“w-CO”), mainly residential but including an industrial user with high water demand (around 40% of the district’s annual consumption) monitored by a smart meter. For w-CO, the inflow data (in L/s) is available at 15-min intervals from January 2022 to July 2024, as well as the repair logs during that period with corresponding repair dates (intervention assignment, start and end of the work). The tree-based unsupervised machine learning algorithm known as Isolation Forest [2] is used to identify anomalies in the dataset, which are cross-referenced with repair records [3,4], mainly indicating the occurrence of leakage or damage to pipes and connections. To enhance anomaly detection at the DMA-level, SCADA system data is integrated with that obtained from remote meter reading. Indeed,



Academic Editors: Patrizia Piro, Bruno Brunone, Federico Roman, Umberto Sanfilippo and Michele Turco

Published: 7 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

the procedure via IF is repeated on an adjusted inflow dataset, derived by subtracting the consumption (in L/h) of the hydro-demanding industrial user.

2. Materials and Methods

Abnormal flow rates are identified using the IF algorithm [2], which is based on the concept that outliers are typically easier to isolate using separating hyperplanes than normal data points. A hyperparameter called contamination fraction (CF) sets the outlier/non-outlier classification threshold. In this study, a range of CFs varying from 0.001 to 0.3 is tested. First, raw sensor flow data is pre-processed to smooth spikes caused by system usage (filter backwash) in w-CO, thus increasing the efficiency of the anomaly detection algorithm with equal CF.

2.1. Isolation Forest

The unsupervised machine learning algorithm builds multiple binary trees by:

1. Drawing a random sample from the data;
2. Selecting a random feature (time of the day, day of the week, month of the year) and a random split value between the min and max of that feature;
3. Recursively splitting until a maximum tree height is reached, or all instances in the node are identical, or only one instance remains.

Once the forest is built, the anomaly score is computed for each data point, based on the average path length required to isolate it across all trees. The Isolation Forest algorithm was selected for this project due to its advantages, including computational efficiency and its ability to handle large volumes of data.

2.2. Metric

The anomalies identified by the IF algorithm are cross-referenced with repair records, mainly indicating the occurrence of leakage or damage to pipes and connections. It is assumed that each repair assignment date corresponds to the date when the anomaly occurred, which marks the start of a corresponding time window, terminating with the completion date of the work. We restrict our analysis to windows with a duration of six days or less, assuming they correlate with more significant events. The definition of a metric allows the relationship between detected anomalies and performed interventions to be quantified, providing a useful tool to evaluate the effectiveness of the algorithm across different contamination values. The metric is computed as follows:

$$M = \frac{\sum_{x=1}^n s_x \times d_x}{na} \quad (1)$$

where:

- s_x = (inverse of) anomaly score assigned by the model to point x ;
- d_x = distance between point x and the first subsequent intervention (if any);
- na = number of anomalies detected for a given CF.

Low values of the metric indicate highly anomalous samples that are temporally close to an intervention.

2.3. Datasets

In this study, anomaly detection is carried out on the following w-CO's flow datasets:

1. 15-min interval inflow rates;
2. Daily minimum flows (DMFs);
3. 1-h interval inflow rates (dataset (1) resampled);

4. 1-h interval inflow rates—industrial user’s consumption.

The most common methods for identifying leaks utilize the concept of minimum night flow (MNF), which recognizes that water usage during night-time hours is less variable compared to the daytime [5]. In this study, the concept of MNF is extended to DMF and anomaly detection is applied to the corresponding dataset (2). Hence, the DMFs can be used as a baseline for comparison with new minimum flow data, with a significant increase indicating a leakage. It is expected that an automated analysis of DMFs could simplify what is currently a periodic analysis carried out by employees of the water utility. The idea behind dataset (4) is to identify anomalies potentially harmful to the system, discarding those attributable to authorized consumption within the DMA.

3. Results and Discussion

The metric is computed for each dataset and CF and the resulting curves are shown in Figure 1. It is clear that the curve representing the metric for daily minima is consistently lower than the others, suggesting that minima are the best candidate for detecting water leakages. It can be observed that for very low contamination values, the model identifies few anomalies, which are strongly correlated with the interventions. As the CF increases, a greater number of anomalies is detected, but with a larger distance from the nearest intervention. To assess the impact of the large user in the DMA, curves (3) and (4) are compared. The results show that the latter remains consistently below the former (at least for $CF < 0.05$), supporting the conclusion that subtracting the industrial user’s consumption from the total inflow to the DMA enhances the performance of leakage detection. This is consistent with what is observed when comparing results at the same CF, namely that anomalies visible in the 15-min interval data may be missed in the hourly resampled data, but re-emerge once the industrial user’s contribution is subtracted.

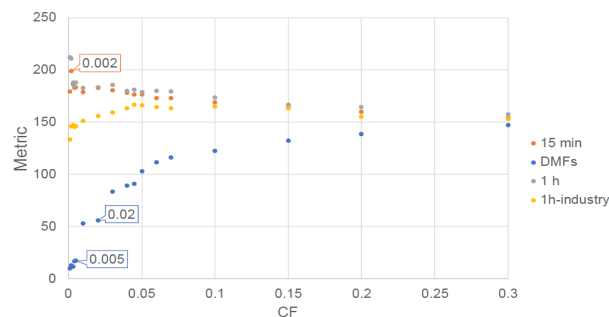


Figure 1. Variation in the metric with contamination fraction for the four datasets.

Results on 15-min interval data show that sensor failures and data transmission errors are easily detected on this dataset as they commonly result in unrealistic plateaus (occasionally with null values) or erratic and spiky trends. In contrast, when using daily minima for the same day and CF, these anomalies are not always detected, since they may be disregarded by the daily minimum value. Regarding DMFs, the IF algorithm is generally able to capture their increase which may be indicative of a leakage, and it flags them as anomalies that often fall within the repair time windows. It is worth noting, however, that gradual increases in flow are rarely detected in the other datasets, whereas sudden changes are consistently and effectively identified.

4. Conclusions

This study has shown that sensor malfunctions and communication anomalies can be effectively detected through the application of the IF algorithm. Regarding burst and leakage events, the automated analysis of DMFs has demonstrated the potential to simplify

tasks that are currently carried out manually by water utility operators. Furthermore, for leakage detection, it appears particularly promising to examine DMFs and inflow time series once they have been cleansed of industrial water consumption. Importantly, the proposed approach is fully compatible with the current configuration of water distribution systems. As such, it offers the possibility of achieving a more time- and cost-efficient identification of anomalies at the DMA-level, thereby enhancing protection against service disruptions and major water losses.

Author Contributions: Conceptualization, C.C. and C.B.; methodology, M.L.; software, L.P.; validation, M.L. and C.C.; formal analysis, C.C. and L.P.; investigation, C.C.; resources, G.N.; data curation, C.C. and L.P.; writing—original draft preparation, C.C.; writing—review and editing, C.C. and C.B.; visualization, C.C. and L.P.; supervision, M.L. and C.B.; project administration, M.L. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: The contribution of Cristiana Bragalli was carried out within the RETURN Extended Partnership and received funding from the EU Next-Generation EU (NRRP, Mission 4, Component 2, Investment 1.3—D.D. 1243 2 August 2022, PE0000005).

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of this data. Data was obtained from Hera S.p.A. and is available with the company's permission.

Conflicts of Interest: Author Giuditta Nicoli was employed by the company Hera S.p.A. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IF	Isolation Forest
DMA	District Metered Area
SCADA	Supervisory Control and Data Acquisition
CF	Contamination Fraction
DMF	Daily Minimum Flow
MNF	Minimum Night Flow

References

1. Wu, Z.Y.; Chew, A.; Meng, X.; Cai, J.; Pok, J.; Kalfarisi, R.; Lai, K.C.; Hew, S.F.; Wong, J.J. Data-driven and model-based framework for smart water grid anomaly detection and localization. *Aqua Water Infrastruct. Ecosyst. Soc.* **2022**, *71*, 31–41. [\[CrossRef\]](#)
2. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422. [\[CrossRef\]](#)
3. McMillan, L.; Fayaz, J.; Varga, L. Flow forecasting for leakage burst prediction in water distribution systems using long short-term memory neural networks and Kalman filtering. *Sustain. Cities Soc.* **2023**, *99*, 104934. [\[CrossRef\]](#)
4. McMillan, L.; Fayaz, J.; Varga, L. Domain-informed variational neural networks and support vector machines based leakage detection framework to augment self-healing in water distribution networks. *Water Res.* **2024**, *249*, 120983. [\[CrossRef\]](#) [\[PubMed\]](#)
5. García, V.J.; Cabrera, E.; Cabrera, E., Jr. The Minimum Night Flow Method Revisited. In Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium, Cincinnati, OH, USA, 27–30 August 2006; pp. 1–18. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.