Combining SCADA and smart meter data to improve anomaly detection at the DMA-level

C. Cincotta[1*], L. Pedroni[2], M. Lombardi[2], G. Nicoli[3], L.C. Zingali[1], C. Bragalli[1]
[1] Dept. of Civil, Chemical, Environmental and Materials Engineering – DICAM, University of Bologna, Bologna, Italy
[2] Dept. of Computer Science and Engineering – DISI, University of Bologna, Bologna, Italy
[3] Hera S.p.A., Modena, Italy
* e-mail: chiara.cincotta@unibo.it

Introduction

In the past decade, increasing adoption of smart sensors in addition to SCADA (Supervisory Control and Data Acquisition) systems has motivated the development of novel data analytics to improve the quality and efficiency of drinking water services (Wu et al. 2022). A vast amount of data, commonly referred to as big data, is collected by this advanced metering infrastructure. A critical issue in utilizing this data for water system management and operation is the prompt detection of anomaly events, which may result from sensor malfunctions, communication failures, new pipe breaks and unauthorized water uses.

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 district metered areas (DMAs), where flow is measured at the inlet and outlet points at regular time-intervals. This study focuses on a specific DMA ("w-CO"), which is mainly residential and where smart water metering has been applied for more than 3 years at around 35% of the users. Furthermore, w-CO includes an industrial user with high water demand which is monitored by multiple smart meters. For w-CO, the net flow data (in l/s) is available at 15-minute intervals from January 2022 to July 2024, as well as the repair logs during that period, with corresponding indication of urgency (deferrable/non-deferrable) and repair dates (intervention assignment, start and end of the work). The tree-based unsupervised machine learning algorithm known as Isolation Forest (Liu et al. 2008) is used to identify anomalies in the dataset, which are cross-referenced with repair records (McMillan et al. 2023, 2024), mainly indicating the occurrence of leakage or damage to pipes and connections.

The main novelty of this study lies in the effort to leverage all available field data to enhance anomaly detection at the DMA-level and this is achieved by integrating SCADA system data with that obtained from remote meter reading. By comparing w-CO net flow data with consumption data of the large industrial user present in the DMA, the significant impact of the latter on the former is highlighted, thus justifying the attempt to apply the Isolation Forest (IF) algorithm to net flow time series cleansed of industrial water consumption.

Materials and methods

Abnormal flow rates are identified using the IF algorithm (Liu et al. 2008), which is based on the concept that outliers are typically easier to isolate using separating hyperplanes than normal data points. In terms of decision trees, this places outliers closer to the root node than expected data points. A hyperparameter called Contamination Fraction (CF) sets the outliers/non-outliers classification threshold. In this study, a range of CFs varying from 0.001 to 0.05 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 CFs. 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. These windows are then labelled as deferrable or non-deferrable and cross-referenced with anomalies identified in the flow dataset. Non-deferrable interventions are supposed to be associated with major (and readily detectable) leakage events. In this

study, anomaly detection is carried out on the following w-CO's flow datasets: (1) 15-minute interval flow rates, (2) daily minimum flow rates, (3) 1-hour interval flow rates obtained from (1), (4) 1-hour interval flow rates calculated as (3) minus hourly consumption values of the industrial user. 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 (García Vicente et al. 2012). In this study, the concept of MNF is extended to daily minimum flow (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 increasing 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 applying IF to dataset (4) is to identify anomalies potentially harmful to the system, discarding those attributable to actual authorized consumption within the DMA. Given the presence of an industrial user with highly variable water demand (l/h), it is decided to subtract its contribution from the net inflow into the district, which is aligned with the large user's dataset, resulting in dataset (3).

For each dataset and CF, anomalies identified through IF are counted and marked with a cross. In this way, it is possible to assess the performance of the algorithm in terms of both the nature of detected anomalies (e.g., a cross within a repair window is assumed to correspond to a leakage event) and the efficiency in identifying all and only the actual anomalies in the datasets.

Results and concluding remarks

It is observed that the algorithm performs well in identifying both extreme outliers and extended periods of unusual flow rates. Sensor failures and data transmission errors are indeed easily detected as they commonly result in unrealistic plateaus (occasionally with null values) or erratic and spiky trends. A striking example of extreme anomalies that affected w-CO and were effectively detected by the IF algorithm, even for the simplest flow series (1) and for very low CFs (e.g., CF = 0.001), is the two floods occurred in the W-CO area in May 2023 due to the breach of a riverbank. As far as burst/leakage events are concerned, some of the unusual flow data points flagged by the algorithm correlate well with repair interventions, while some repair dates are observed to be away from the outlier data even for high values of contamination, probably due to the limited magnitude of the burst/leakage event. The IF algorithm tends to capture increases in the DMFs, identifying them as anomalies that frequently fall within the repair time windows. The gradual increase in flow is almost never detected in (1), (3) and (4), whereas sudden changes are effectively identified. It can therefore be concluded that the anomaly detection of flows at 15-minute or hourly intervals and DMFs are somehow complementary. Moreover, with the aim of detecting leaks, it seems promising to consider DMFs derived from (4), as they are not influenced by the industry.

The proposed approach is compatible with the current set-up of water distribution systems managed by most water utilities. It may facilitate a time- and cost-efficient identification of anomalies at the DMA-level and better protect it from disruptions and major water losses, which is an important step towards smart and sustainable urban water management.

References

García Vicente J, Cabrera E, Cabrera Jr E (2012) A significant and interesting scientific report. Journal of Scientific Reports 205(4): 83-97. https://doi.org/10.1061/40941(247)35

Liu FT, Ting KM, Zhou Z-H (2008) Isolation Forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp 413-422. https://doi.org/10.1109/ICDM.2008.17

McMillan L, Fayaz J, Varga L (2023) Flow forecasting for leakage burst prediction in water distribution systems using long short-term memory neural networks and Kalman filtering. Sustainable Cities and Society 99: 104934. https://doi.org/10.1016/j.scs.2023.104934

McMillan L, Fayaz J, Varga L (2024) Domain-informed variational neural networks and support vector machines based leakage detection framework to augment self-healing in water distribution networks. Water Research 249: 120983. https://doi.org/10.1016/j.watres.2023.120983

Wu ZY, Chew A, Meng X, Cai J, Pok J, Kalfarisi R, Lai KC, Hew SF, Wong JJ (2022) Data-driven and model-based framework for smart water grid anomaly detection and localization. AQUA - Water Infrastructure, Ecosystems and Society 71(1): 31–41. https://doi.org/10.2166/aqua.2021.091