A modern approach for leak detection in water distribution systems

Alexandru Predescu
Automatic Control and Computers
University POLITEHNICA of Bucharest
mircea.predescu@stud.acs.upb.ro

Mariana Mocanu
Automatic Control and Computers
University POLITEHNICA of Bucharest
mariana.mocanu@cs.pub.ro

Ciprian Lupu
Automatic Control and Computers
University POLITEHNICA of Bucharest
ciprian.lupu@acse.pub.ro

***Abstract*—Water distribution is arguably the most important factor in modern times. The quality of service is nonetheless taken for granted while requiring regular maintenance and timely repairs. Therefore, there is an increased demand in efficient operation requiring modern solutions to age-old problems. Multiple theoretical and practical approaches exist for leak detection. In this paper, a model-based approach is extended with modern concepts from the field of Machine Learning for improving leak detection accuracy in water distribution systems.**

***Index Terms*—Leak Detection, Model Simulation, Machine Learning, Water Distribution System**

I. Introduction

Water distribution systems are an essential part of modern world while being nonetheless taken for granted. The hidden infrastructure makes it difficult to detect signs that can predict potential problems over long periods of time, such as possible contamination of water, an increased energy usage and environmental damage [1].

The quality of service in utility networks, and the particular case of water distribution systems, is closely related to the quality of the infrastructure and efficient operation.

As new technologies emerging from fields such as ICT (Information and Communications Technology) are being regarded as possible solutions to the problem of efficient resource management in present infrastructure, there is a great deal of research on the multiple theoretical and practical approaches related to the problem of leak detection in mass transfer systems [1], [2], [3], [4], [5]. These ICT systems collect data for monitoring and control purpose and represent an important layer of infrastructure in water distribution systems [6].

There are several traditional approaches for monitoring the state of the system and for detection of problems:

A. Hardware solutions, such as using acoustic sensors, gas detectors, negative pressure detectors and infrared thermography, as presented in many literature approaches

B. Software solutions, using modeling and simulation of flow and pressure and real-time event monitoring such as SCADA (Supervisory Control And Data Acquisition) systems

In the case of hardware solutions, the advantage is given by the accurate location of identified leaks. It is however an expensive solution and not very time efficient.

The modeling solutions take into consideration the *Conservation of Mass* that assumes measuring the input and output from the system and setting a threshold on the difference to signal a leak. It is possible to measure the change in pressure/flow, with some considerations regarding the cost of the required sensors and the accuracy as described in [7]. In a model-based approach, real-time measurements are compared with the expected values from a hydraulic model to highlight possible discrepancies.

Another possibility is the real-time evaluation of the loss of pressure/flow in measurement nodes. This method is described in [8].

Recent developments in the field of ICT, allow for an improved response time for leak detection, considering the mass adoption of smart meters. The estimation of water demand from smart meter data allows for efficient operation and anomaly detection, as described in [9]. While the accuracy is closely related to the data acquisition solution of choice, it is usually not possible to estimate the exact location of leaks.

Nonetheless, the aforementioned methods can be complementary, in the sense that software solutions are recommended for providing quick detection and a rough estimation of the location of leaks, while the hardware solutions can be used for accurate location. The final objective is to fix the problem in the shortest amount of time and with minimal costs. This still assumes the role of the operator, that has to take decisions based on the information that is provided by the monitoring system.

The integration of Machine Learning in a broad range of applications provides a foundation for a new paradigm, where the cognitive process is shifted from the operator towards AI (Artificial Intelligence). This allows for improved data processing capabilities, by extracting relevant information from vast amounts of data [10], [11]. In this paper, we can identify a modern method for leak detection that integrates industry standard solutions with Machine Learning algorithms (**C.**).

II. Related work

The traditional approaches that we consider in this paper are described in numerous papers and are being used and standardized in the industry. In the case of leak detection using modeling techniques, there are methods based on steady-state

operation and transient effects. This is also part of a regulatory framework, with requirements for the pipelines transporting liquid and gas in Germany.

There are several possibilities regarding the choice of internal sensors that can detect a leak, such as pressure sensors, flow sensors and temperature sensors. In the case of pressure sensors, that are more commonly installed and generally less expensive, a leak is regarded as a pressure drop Δ_p. When using flow sensors, the leak can be directly measured and the accuracy is usually higher, as described in [7]. These methods work best under steady-state conditions e.g. during the night, when the consumer demand is low and relatively constant. There are two types of leak signatures that can be identified: sudden leaks and gradual leaks [5]. A leak signature analysis is important to predict false alarms while maintaining accuracy and a fast response time.

On the other hand, Machine Learning is currently being used in multiple areas and the concept of learning from the available data while expanding the knowledge base that is used for cognitive functions is a kind of self-regulating process. The adaptability of such methods to various fields allows for a common set of algorithms that can be used to extract relevant information from the available data.

According to [12] it is possible to expand the knowledge base by unsupervised learning, using the *k*-means algorithm. Moreover, real-time data processing capabilities are demonstrated in [13]. Related work in the field of leakage management, using a Machine Learning based approach, is presented in [14], [15]. The identification of leaks based on the transient characteristic is also described in [16].

In this paper, we combine the traditional methods of dynamic modeling with the more recent solutions from the field of Machine Learning, in a proactive approach to water network management.

III. Proposed solution

We propose the fusion between the traditional model based approach for leak detection and a modern algorithm for leak signature classification. The real-time monitoring solution is based on a SCADA system and a storage solution that records the measured and simulated parameters. The simulated parameters can be provided by a hydraulic model or by a previously identified pattern, such as the result of a clustering algorithm. The modern approach that we propose for leak detection consists in real-time clustering and classification of the deviation pattern that is obtained when comparing the measured and simulated parameters (or when comparing the normal conditions to the current measurements).

In Fig. 1 is shown an overview of the proposed solution. There are two alternative methods shown. The first method is to use a hydraulic model to generate a reference model for measured data. The second method is to use the data from a clustering algorithm as a reference, which can reduce the complexity added by using a calibrated hydraulic model.

The data that we used in this paper is provided by real measurements using smart meters that were installed in multiple locations in Italy. The data is represented as time series for individual consumers, that show the daily measurements with a sampling time of 1 hour over a time-frame of multiple months. As the data is not correlated with the network configuration that we use in this paper, we use the measurement data that we assign to a single consumer node and we estimate the flow for the other nodes using a hydraulic model.

The model of the water distribution system is represented by an undirected graph and the fundamental laws that define the water flow through a pipe are used to calculate the state of the network based on the available data.

The first law is the mass conservation law that states that the input flow to a node is equal to the output flow:

$$\sum_j q_{ij} - \sum_j d_{ij} = 0,\ i = 1..n \tag{1}$$

The second law is the equivalent of Ohm's law for laminar flow and gives the value of the flow for a network segment between two adjacent nodes:

$$q_{ij} = \frac{h_i - h_j}{R_{ij}} \tag{2}$$

$$R_{ij} = \frac{8\eta l_{ij}}{\pi r_{ij}^2} \tag{3}$$

The dynamic model is simulated using a first order filter with the parameters according to the physical characteristics of the network segment:

$$G_{ij}(s) = \frac{1}{T_{ij}s + 1} \tag{4}$$

We used the following notations:
q_{ij} - input flow from node i to j
d_{ij} - output flow from node i to j
h_i - head (pressure) in node i
R_{ij} - resistance to flow in the pipe between node i and j
η - fluid viscosity
l_{ij} - length of the pipe between node i and j
r_{ij} - radius of the pipe between node i and j
$G_{ij}(s)$ - transfer function of the filter for the pipe between node i and j
T_{ij} - filter time constant
s - Laplace transform variable

We propose an extension of the algorithm to simulate a first order dynamic model. The unknown variables that are calculated by the static model are filtered using a first order transfer function, with a time constant proportional to the dimensions of the pipe (length, diameter) as in eq. (3).

We input the measured data for a given node to the algorithm and the instantaneous results obtained for a single sample are being fed back to the subsequent iterations, after being filtered by the corresponding pipe model.

In the same way as there are different types of consumer patterns (e.g. residential, commercial, industrial), the fault

conditions can be classified and correlated with these patterns [17].

In the following paragraphs, we define the following notations for data sets:

D1 - the original measurement data set, $\mathbb{R}^{N_N \cdot N_D \cdot 24}$

D2 - the altered measurement data set, $\mathbb{R}^{N_N \cdot N_D \cdot 24}$

SM1 - the simulated measurements using the original data set, $\mathbb{R}^{N_N \cdot N_D \cdot 24}$

SM2 - the simulated measurements using the altered data set, $\mathbb{R}^{N_N \cdot N_D \cdot 24}$

DSM - the difference between simulated measurements using SM1 and SM2, $\mathbb{R}^{N_N \cdot N_D \cdot 24}$

DCSM - the clusters obtained using the difference data set, $\mathbb{R}^{N_N \cdot N_C \cdot 24}$

DFC - the second stage centroids that represent the average leak patterns for the entire network, $\mathbb{R}^{N_{DC} \cdot 24}$

where:

N_N - number of measurement nodes

N_D - number of days

N_C - number of centroids for clustering

N_{DC} - number of centroids for second stage clustering

From the perspective of a single node, we consider the original data (D1) and then we add a simulated leak to the data set (D2) in order to test different leak scenarios. The data set for a single node consists in multiple 24-samples time-series for each day, so it is possible to evaluate the effects of the leak over a broad range of measurements. Therefore, we subtract the simulated time-series and the result ($\mathbb{R}^{N_D \cdot 24}$) is input to a clustering algorithm to extract the leak patterns ($\mathbb{R}^{N_C \cdot 24}$).

In order to obtain an extensive classification of the leak patterns, we consider extending this method to measure the effects of a leak on multiple nodes in the network. Therefore, to estimate the data for the adjacent nodes, we use a simulation model that is described in [7]. The experimental setup is presented in Fig. 2 and is described in the following paragraphs:

We define a constant supply pressure at the input node and output nodes. Then, the measurements from the original data set (D1) are assigned to a consumer node and the model returns the estimated values for the remaining nodes. Therefore, the model simulation of steady-state conditions is run for each sample in the time-series and the result is a data set with the simulated parameters (SM1).

This entire simulation is repeated using the altered data set (D2) and we obtain the second set of simulated parameters (SM2). The difference data set (DSM) is calculated by subtracting the two data sets. Then the clustering algorithm reveals the common patterns for the test scenario (DCSM) calculated for each individual node. This allows for a cleaner representation of the data, when compared to the large amount of original time-series. For a general overview, a second-stage clustering algorithm aggregates the clusters to define the general patterns in the network for the analyzed leak scenario.

The difference data set (DSM) can be further used to generate a fault sensitivity matrix that represents the overall (average) sensitivity of the network during the analyzed test scenario, as an extension to the static fault sensitivity analysis described in [7]. The matrix can be used to show a dynamic overview of the network sensitivity, as well as an average sensitivity of the dynamic model simulation.

We define the following test cases to validate the solution:

A. The constant leak scenario is simulated by adding a step function (constant value) to the measurements within a specific time frame (we considered the time frame 12-18)

B. The gradual leak scenario is simulated by adding a ramp function (linear increasing value) to the measurements

C. The sudden leak scenario is simulated by adding an impulse function (local value) to the measurements at specific points in time. This scenario is only relevant to pressure measurements though, which are not available in the data set.

This method uses the data from each measurement node to find a general pattern for a given leak scenario, that can be used to evaluate the subsequent leak profiles. This benefits from accumulating data, by expanding the knowledge base in the sense of a more general representation for a given leak scenario, that can include multiple other particular representations, which are actually part of the same class.

A. Python implementation

The Python language is used for the Machine Learning and Data Processing algorithms. The data is stored in CSV files, that represent the associated data for a measurement node. The parsing returns the data to a numpy array, that is useful for processing, providing a standard representation libraries that are designed for solutions in the field of data science. The most common Machine Learning algorithms are implemented in the *scikit-learn* package, that uses scipy, as well as numpy and matplotlib [18].

For unsupervised learning that we require for pattern identification in the case of both the consumer demand profile and the leak signature, we use the *k*-means algorithm from the *sklearn.cluster* package that implements the automatic grouping of similar data into sets. The method *fit* implements the *k*-means clustering of the input data that in different shapes. The algorithm has three steps, the first is choosing the initial centroids that can be a sample from the data set. The following two steps are used iteratively, until a certain stop condition is met. The first assigns samples to the nearest centroid and the second recalculates the centroids by averaging the previously assigned samples for each centroid. The results, for example the cluster centroids, can be extracted from the output.

This way, we apply the algorithm to the original data set and to the altered data set and we obtain two series of centroids for each measurement node. Then, we calculate the difference between the corresponding centroids for each particular node and we apply the algorithm once again to find the patterns for each test scenario.

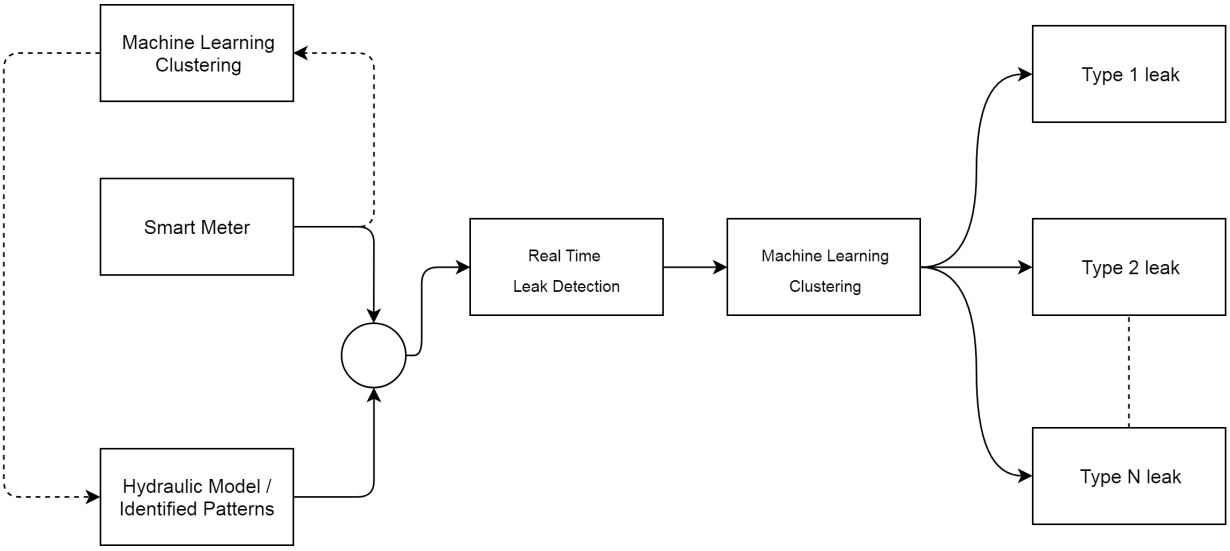


Fig. 1: Solution overview

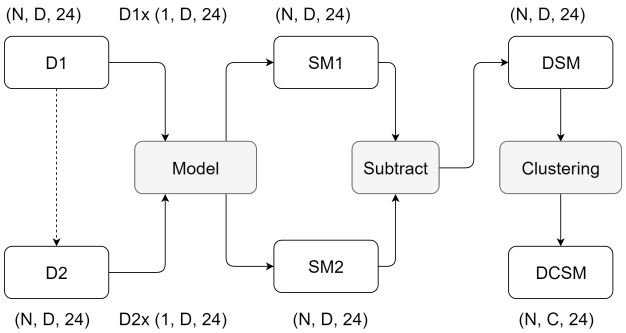


Fig. 2: Experimental setup

IV. Results

The purpose of the experiments is to validate the method for leak pattern identification in a water distribution system. The figures represent the time-series representation of the leak patterns during a 24-hour window. This dimension is used as the original data set is comprised of daily time series for each node. The values of the centroids represent the average flow for the particular pattern. In the same way are represented both the consumer patterns and the simulated leak scenarios.

The patterns for the simulated data using the original data set (normal conditions) are shown in Fig. 3 for the first stage clustering of individual consumers. The second stage clustering is applied on the aggregated first stage clustering results from each consumer node, using a fixed number of 3 clusters. The result of the second stage clustering is shown in Fig. 4(a), 5(a).

We present the patterns of simulated data for the proposed test cases for a better overview of the effects of a leak on the measured parameters in the network. The difference data between the normal conditions and the simulated test case is input to the clustering algorithm that extracts the leak patterns and provides a clear representation of the particular scenario.

The first scenario (A) is represented in Fig. 4 and the second scenario (B) is represented in Fig. 5. In the (a) figures, the results are obtained using the second-stage clustering of the consumer data, and in the (b) figures, the results are obtained using the first-stage clustering of the difference data for the corresponding leak scenario (DSM). The second-stage clustering of the difference data (DFC) is presented in Fig. 6.

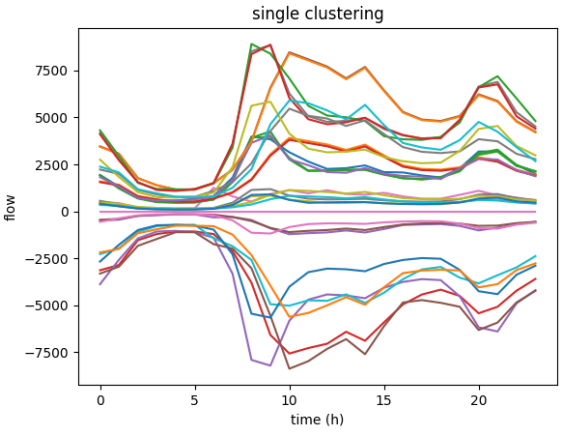
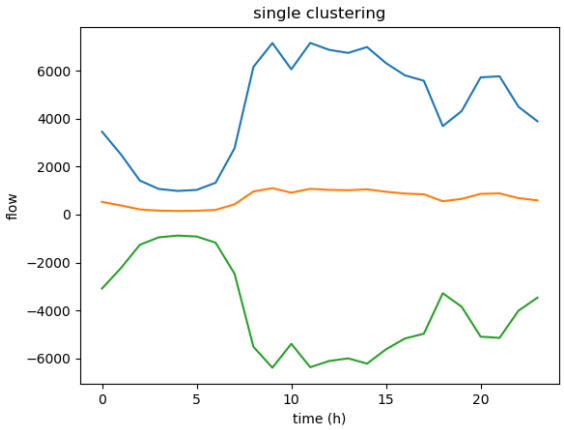
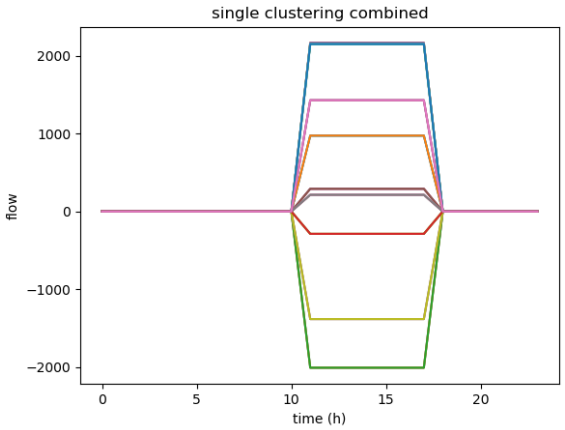


Fig. 3: First-stage clustering with normal conditions

The results show an accurate representation of each test case, with the cluster centroids obtained from the difference data set. The shape of the centroids is consistent and only the absolute value is different, as it represents the deviation for a particular node in the network. As the first model simulates only the steady-state conditions, the transient effect is not emphasized in this simulation. Using a simple first-order dynamic model that takes into consideration the length of the pipes, we obtain the centroids in Fig. 6.

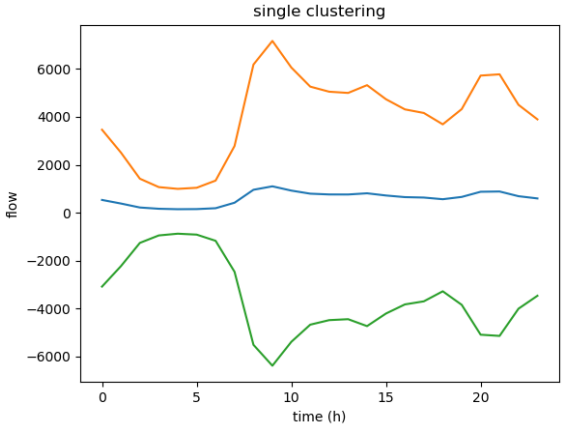


(a) Cluster centroids from simulated data

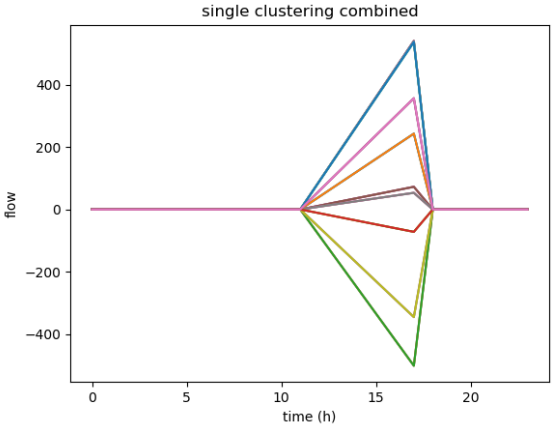


(b) Cluster centroids from difference data

Fig. 4: Clustering of simulated data with test case A

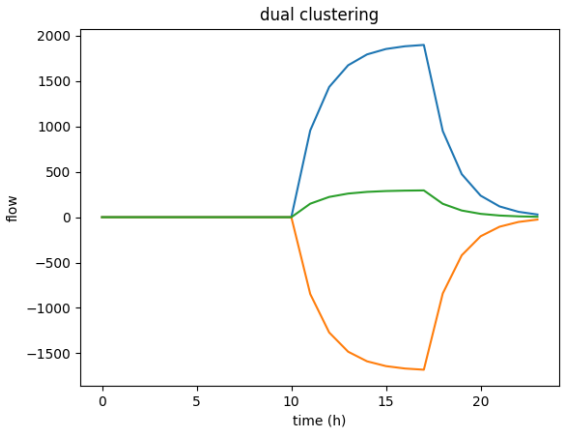


(a) Cluster centroids from simulated data

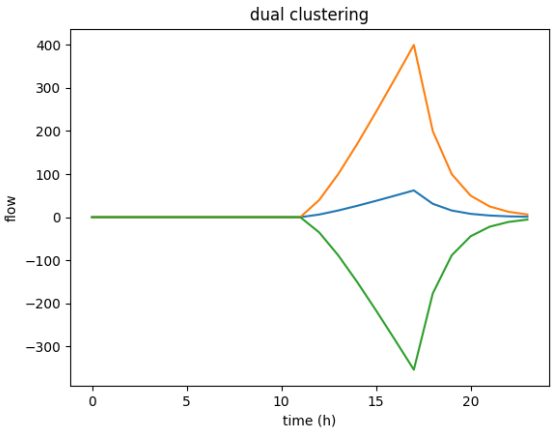


(b) Cluster centroids from difference data

Fig. 5: Clustering of simulated data with test case B



(a) Test case A difference clusters



(b) Test case B difference clusters

Fig. 6: Leak scenarios. Second-stage clustering with dynamic model

V. Conclusion

In this paper, the modeling approach and hardware solutions are being referred as traditional methods for leak detection in water distribution systems. The methods are nonetheless valuable for the industry and provide a foundation for higher-level cognitive solutions that emerge from the field of Machine Learning. The highly technical problem of evaluating the cause of the leak using specialized sensors and standard software algorithms is presented in this paper from a different perspective that arises from unsupervised learning algorithms. The shift from the high precision requirements of a real-time sensor data analysis to the Big Data paradigm can be an important step for increasing the level of mass adoption for smart infrastructure, by reducing the costs associated to high performance hardware solutions.

The proposed solution is based on the unsupervised clustering of data from smart meters, which allows for a more accurate evaluation of the root cause of the leak, as this can range from localized damage (e.g. construction works accidents) to gradual leaks that can be caused by aging infrastructure. Each of the aforementioned causes can be represented as a specific pattern that is identified by the modern solutions.

In the case of enhancing the traditional methods with modern solutions, the hydraulic model can provide a reference to the clustering algorithm, especially in the case where the data is not available and has to be estimated. On the other hand, the clustering algorithm can be used for calibrating the hydraulic model. The clustering algorithm provides a foundation for more advanced solutions, such as anomaly detection, and can be adapted to any situation that requires extracting relevant information from vast amounts of data.

Acknowledgement

We are thankful to the PN III Program P3 European and International Cooperation, UEFISCDI, that supported the research activity and part of the presentation in conference, as well as to the H2020 Twinning Program, that partially supported the publication under the 690900 project - Data4Water

References

[1] S. Oven, "Leak detection in pipelines by the use of state and parameter estimation, master thesis," Norwegian University of Science and Technology, Department of Engineering Cybernetics, Tech. Rep., January 2014.

[2] C. Lupu, D. Chirita, S. Iftimie, and R. Miclaus, "Consideration on leak/fault detection system in mass transfer networks," in *Energy Procedia*, vol. 112, March 2017, pp. 58–66.

[3] R. Isermann, "Process fault detection based on modeling and estimation methods—a survey," in *Automatica*, vol. 20, no. 4, July 1984, pp. 384–404.

[4] N.C. Turner, "Hardware and software techniques for pipeline integrity and leak detection monitoring." Society of Petroleum Engineers, 1991.

[5] G. Geiger, *Principles of Leak Detection*, 2012.

[6] M. Umar and W. Uhl, "Integrative review of decentralized and local water management concepts as part of smart cities (lowasmart)," Norsk institutt for vannforskning, Tech. Rep., 2016.

[7] A. Predescu, M. Mocanu, and C. Lupu, "Modeling the effects of leaks on measured parameters in a water distribution system," in *2017 21st International Conference on Control Systems and Computer Science (CSCS)*, May 2017, pp. 585–590.

[8] ——, "Real time implementation of iot structure for pumping stations in a water distribution system," in *2017 21st International Conference on System Theory, Control and Computing (ICSTCC)*, Oct 2017, pp. 529–534.

[9] D. García, D. Vidal, J. Quevedo, V. Puig, and J. Saludes, "Water demand estimation and outlier detection from smart meter data using classification and big data methods," February 2015.

[10] M. A. Alsheikh, S. Lin, D. Niyato, and H. P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1996–2018, Fourthquarter 2014.

[11] F. Iglesias and W. Kastner, "Analysis of similarity measures in times series clustering for the discovery of building energy patterns," *Energies*, vol. 6, no. 2, pp. 579–597, January 2013.

[12] A. L. N. Fred and A. K. Jain, "Data clustering using evidence accumulation," in *Object recognition supported by user interaction for service robots*, vol. 4, 2002, pp. 276–280 vol.4.

[13] A. Predescu, C. Negru, M. Mocanu, and C. Lupu, "Real-time clustering for priority evaluation in a water distribution system," in *AQTR'18*, 2018.

[14] A. Candelieri, D. Conti, and F. Archetti, "Improving analytics in urban water management: A spectral clustering-based approach for leakage localization," *Procedia - Social and Behavioral Sciences*, vol. 108, pp. 235 – 248, 2014, operational Research for Development, Sustainability and Local Economies. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042813054748

[15] A. Candelieri, D. Soldi, D. Conti, and F. Archetti, "Analytical leakages localization in water distribution networks through spectral clustering and support vector machines. the icewater approach," vol. 89, pp. 1080–1088, 12 2014.

[16] L. Ferrandez-Gamot, P. Busson, J. Blesa, S. Tornil-Sin, V. Puig, E. Duviella, and A. Soldevila, "Leak localization in water distribution networks using pressure residuals and classifiers," *IFAC-PapersOnLine*, vol. 48, no. 21, pp. 220 – 225, 2015, 9th IFAC Symposium on Fault Detection, Supervision andSafety for Technical Processes SAFEPROCESS 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896315016602

[17] D. García, D. Gonzalez, J. Quevedo, V. Puig, and J. Saludes, "Clustering and classification of aggregated smart meter data to better understand how demand patterns relate to customer type," Universitat Politècnica de Catalunya (UPC), Tech. Rep., 2015.

[18] scikit-learn developers (BSD License). Selecting the number of clusters with silhouette analysis on kmeans clustering. [Online]. Available: http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html