A multistage approach to detect and isolate multiple leakages in district metering areas in water distribution systems

Xiaoting Wang[1], Junyu Li[2], Xipeng Yu[3], Ziqing Ma[4], Yujun Huang[5]
[1,2,3,4,5] Smart Water Research Center, School of Environment, Tsinghua University, Beijing, China
[1] *wang-xt15@mails.tsinghua.edu.cn;*

Keywords: Data-driven approach, flow and pressure estimation, leakage detection, leakage localization, time series decomposition

ABSTRACT

Hydraulic accidents or abnormal situations, also known as leakages, cause not only water losses, but also service interruptions and other negative effects [1]. In order to facilitate the rapid response of water utilities and to reduce water losses caused by undiscovered leakages, a timely detection method is required. Built on data-driven method and hydraulic modelling, this study proposes a multistage approach to solve the leakage detection and localization of the battle problem in L-town.

The proposed approach is shown in fig.1, which consists of three stages:

i) estimation of flow and pressure;
ii) identification of abrupt and incipient leakages; and
iii) isolation of pipes with leakages.

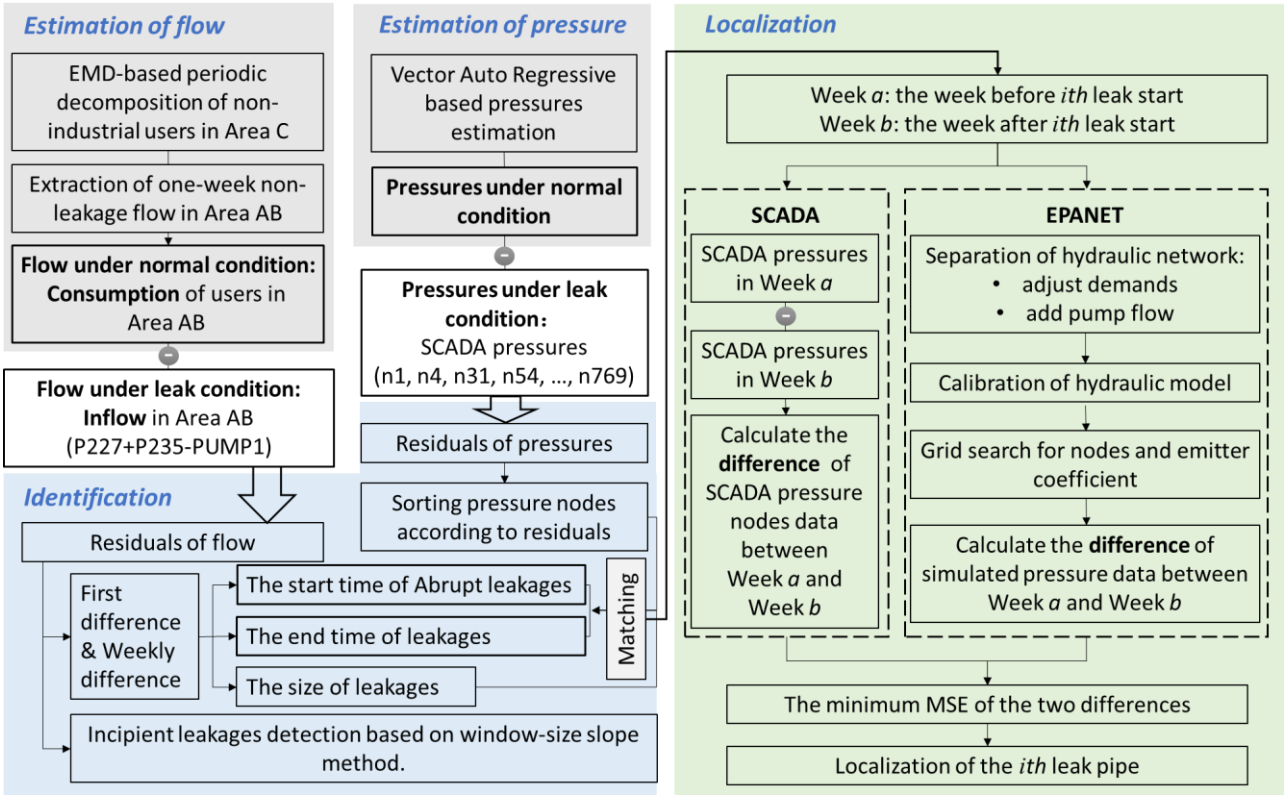

Figure 1. Flow chart of the multistage approach

1. Estimation of flow and pressure

As described in the L-town problem statement, the water utility has set up pressure and flow sensors at a few predefined locations in the network. Time series data collected from these sensors are provided. The variations in the pressure/flow data are caused by daily, weekly and seasonal water usage patterns, leakages and other uncertain factors. To facilitate the identification of leak-induced data from the time series, flow and pressure caused by the users, which are called flow and pressures under normal condition, are estimated by data-driven time series methods.

1.1 Flow

According to the layout of the L-town water distribution network and the locations of the sensors, the network is divided into two relatively independent areas. Area C has sensors for total inflow, and most nodes in the area have water meters (AMRs) at the same time, which is regarded as one zone. Area A and B are treated as one zone, which are called area AB, because there is no separate sensor measuring the inflow to area B. We calculate the flow in area C and area AB separately.

For area C, the difference between the total inflow and total water consumption can reflect the leak information. The total inflow is calculated via the water level in the tank and the flow data of the pump that delivers water to the Tank. It should be noted that the level data of the tank has not changed before and after some time point, and the pump at this time also stopped running. This seems to suggest no water usage in the area, yet it contradicts with the data from the AMRs which show evident water consumption in that period. This implies inaccuracy in the level data of the tower. To minimize its adverse impact on the detection of leakage events in area C, the total inflow and total demand in area C are calculated at intervals of 30 minutes.

The total inflow to the AB area, is recorded by three flow sensors . The actual water consumption of users in the AB area needs to be estimated. Considering the relatively complete monitoring of junction consumption behavior in area C, this can be used as a basis for the prediction of total demand in area AB. Specifically, the changing trend of users' water usage mode in area C during the year should be similar to that in area AB. Therefore, the annual trend in the data in area C can be transplanted to the prediction of water demand in area AB. The water consumption of area AB can be predicted by combining with the simulation in one week of the hydraulic model. The main problem here is to analyze the trend of demand mode in area C.

We assume that the demand in area C is composed of trend items (changes within the year), periodic items (changes within the week), and random items (by stochastic factors), that is

$$Y(\mathrm{t}) = S(\mathrm{t}) \times T(\mathrm{t}) \times R(\mathrm{t}) \qquad (1)$$

where Y(t) represents the total AMR data in the C area, S(t) represents the trend term, T(t) represents the periodic term, and R(t) represents the random term in the data.

To extract the time series items, empirical mode decomposition (EMD) has been employed in this study. EMD can identify the periodic component of time-series data and the trend items of data, it can decompose the original signal via subtracting the intrinsic modal function (IMF) successively, which is the mean of the upper and lower envelopes of the signal to be decomposed [2]. The trend items in demand mode can be represented by one of the IMFs. The trend item in the AMR data is stripped through the multiplication model (Equation 1). Hence, the AMR data needs to be transformed. The logarithmic transformation is used before EMD in this study. Then the trend item is extracted from AMR data of region C. This trend item represents the annual change in water consumption in area C (or area AB). In summary, the estimated value of water consumption in the AB area can be calculated by the following method

$$Demand_{AB,predicted} = S(t) \sum_{i \in N_{AB}} q_i(t) \qquad (2)$$

where $q_i(t)$ represents the demand at node i, which is obtained through the hydraulic model. N_{AB} represent the node set of region AB.

Note that areas with industrial users do not follow the same consumption pattern of the others, hence the AMR data does not include the records of the industrial junctions, which consume huge amounts of water and have high uncertainty. These characteristics will affect the similarity of the two regional (area AB and area C) trend items.

1.2 Pressure

The time series of pressure sensors are influenced by both normal water usages and water leakages. In order to select a suitable forecasting model for normal water consumption, the monitoring data is comprehensively analyzed first [3]. The input data, which is a high dimensional time series data, exhibits strong spatial and temporal correlation [4]. Vector autoregressive model (VAR) has proven to be efficient in many multi-dimensional time series forecasting tasks. VAR model regressively uses historical information and white noise to give the prediction of the current value. In

the meanwhile, the effects of exogenous series are considered as well. As such, we use VAR model to predict the normal pattern of the pressure series. In the training phase, we use a machine learning model trained only with the normal data. Then in the test phase, the well-trained model is applied on the whole dataset and produces the normal pattern of the pressure series. After that, we use the residual between the observed value and the true value in the detection of abnormal value. The residual between the estimated series and the observed series indicates the abnormal behavior of the pressure series.

2. Identification

Flow and pressure are estimated though appropriate prediction, the estimated values are considered as the aggregated user consumption under normal conditions. When a leakage occurs, the abnormal flow and pressure (observed from SCADA) differ from the estimated value, as shown in fig.2. For the flow of area C, data observed from the pump sensor and the tank level meter are used to calculate the total inflow of area C, and the sum of AMR demands gives the aggregate user consumptions. The difference between total inflow and user consumptions is used to identify leakages as residuals in area C.

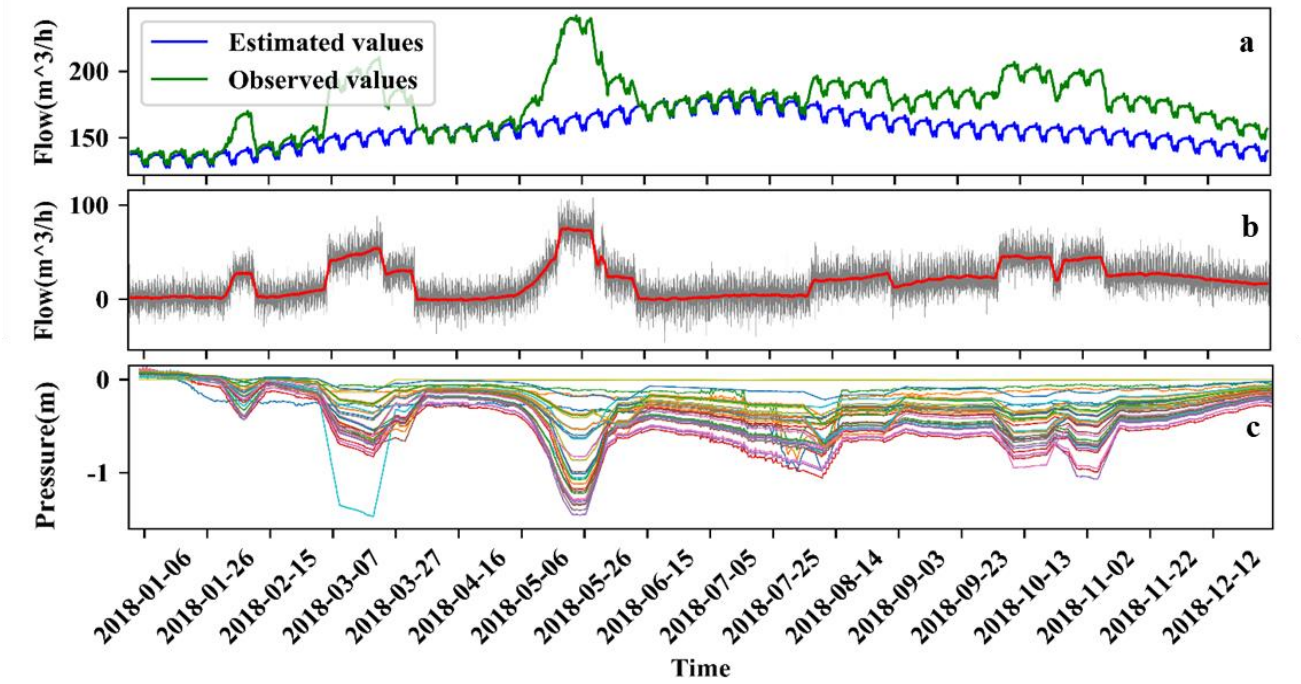

Figure 2. Schematic diagram of leakage identification in Area AB. (a) Estimation of the flow; (b) Residual between estimated and observed flow; (c) Residuals between estimated and observed pressures.

In the pipe networks provided, the leakages are divided into two types: abrupt leakage, incipient leakage. Corresponding to residual data, the pattern of residual series contains constant (no leak), abrupt increase (abrupt leakage), incipient increase (incipient leakage) and abrupt decrease (repaired leakage, the end time of leakage) . Different methods are proposed to deal with leakages of different types. Statistical Process Control (SPC) methods are proposed to solve this leakage detection mission.

As shown in fig. 2b, although the residuals series is calculated by the difference between observed and estimated series, it is still a time series with irregular fluctuations. In order to reduce the nonstationary of residual series, first difference and weekly difference are calculated to make the data smooth, as shown in fig. 3. The abrupt increase and decrease are easily identified through difference transformation, three-sigma SPC limit is used to identify the sudden rise and incline of the flow curve [5]. Therefore, the start time (abrupt increase), the end time of leakages (abrupt decrease), and the leak size (the abrupt value of increase or decrease) are obtained.

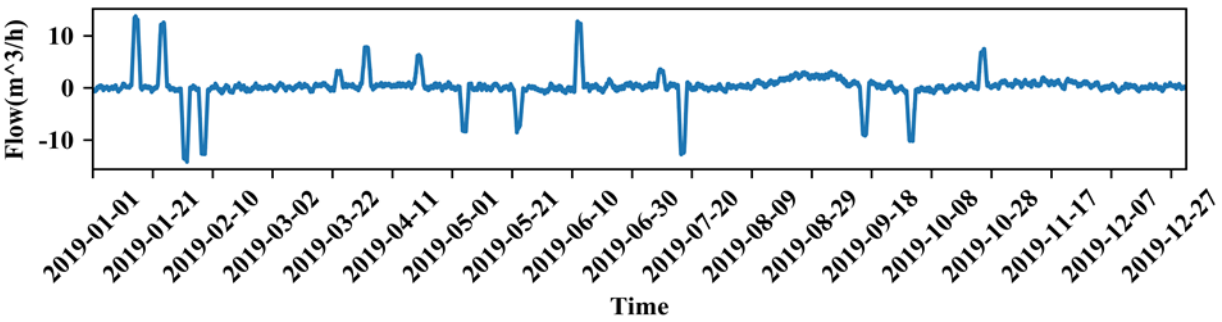

Figure 3. Weekly difference of flow residuals in area AB.

There is one more thing to figure out, how does these start time match the end time. The equality of the variation of the flow at the start and end of the leakage is used to match the start time and the end time of a leakage. A complementing method to address this problem is to sort the pressure nodes residuals, as shown in fig. 4. A sudden increase in pressure will lead to an increase in the residual between the current value and its moving average. The variation of the pressure at the neighboring nodes will be higher than more distant nodes. The birth and fix of a certain leakage will cause similar patterns for the neighbor nodes and the distant nodes. By observing the pattern at each start time and end time, a matching could be done.

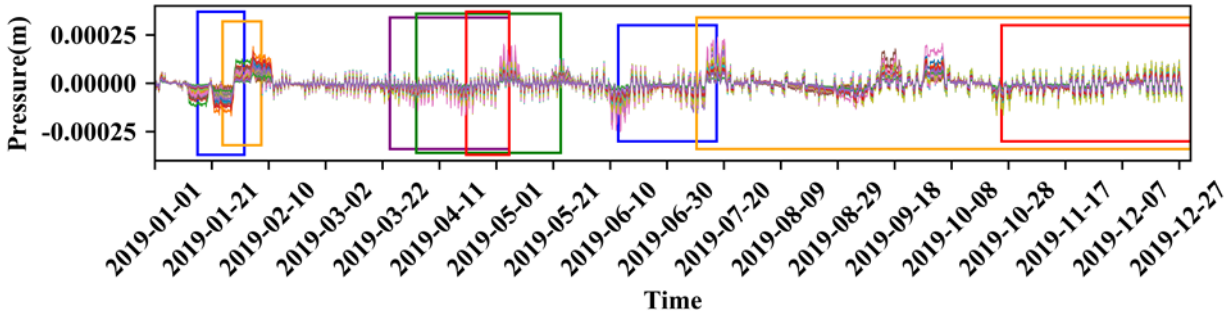

Figure 4. Variation of the pressure for the matching of the start time and end time.

For incipient leakages, a window-size based SPC slope method is employed to identify the incipient increase in flow residuals. If without leakages, the curve of residuals is likely to be smooth and constant. The pattern of flow residuals consists of many sub-patterns caused by different leakages. After the identification of abrupt leakages, the pattern caused by abrupt leakages could be removed from the curve of residuals. The rest of the curve only contains the sub-pattern caused by the incipient leakages. Their slopes could be used to distinguish different unfixed incipient leakages. Each incipient leakage has an identical and constant slope. In this way, the start time of each incipient leakage could be identified once the slope of the curve changes. We use a slide window of length T to go through the curve. As it moves along the time axis, it calculates the slope in the window. A threshold is set for the identification of the incipient leakage. As shown in Table 1, all the unfixed incipient leakage could be identified with the above mentioned method. However, the peak of each unfixed incipient leakages is hard to be determined because of the overlap of different incipient leakage. Therefore, an approximate peak time is given for each incipient leakage.

Table 1. Identified slopes of the incipient leakages in area AB

Start time	Slope(m^3/(h*5min))	Peak time	Leak type
2018/09/10 00:00	0.000508	2018/11/01 21:30	Incipient leakage
2019/03/28 14:15	0.0014	2019/04/07 14:15	Incipient leakage
2019/06/14 15:10	0.00072	2019/10/22 15:10	Incipient leakage
2019/10/25 22:20	0.0015	2019/12/09 22:20	Incipient leakage
2019/11/09 01:10	0.0017	2019/11/29 01:10	Incipient leakage

3. Localization

Due to the fact that the boundaries of the three regions of the L-Town network are obvious, it is easy to build hydraulic models of area C as it has a large number of AMRs, so we split the L-Town

network into two parts, area C network and area AB network, to locate leakages separately. We first delete the pump connecting area AB and area C, so that the original L-Town network is divided. For area AB network, we add a node at the position of the pump with a demand pattern based on pump flow SCADA data, and we replace the tank with a reservoir which is assigned a head pattern using the observed tank level data in area C. Finally, we build two separated networks (areas AB and C) to locate leakages with the same hydraulic calculation results as the L-Town network.

The hydraulic model of L-town built on EPANET tool is employed to simulate nodal pressure. The pressure sensors' data without leakages in 2018 is used to calibrate two networks by adjusting roughness of pipes to minimize the mean squared error (MSE) of observed and simulated pressure values.

For area AB, it is difficult to build an accurate hydraulic model as the network has no AMRs, so we use the idea of control variates to localize each leakage in this area. Firstly, we find two weeks when the leakage had occurred and the leakage had not occurred based on the start and end time of the leakage which is calculated before. Secondly, making a difference between the observed pressure data at the corresponding time in the two weeks. We assume that the leakage to be localized is the only leakage between the two weeks, so the difference could reflect the effect of this leakage on pressure sensors in the network. Thirdly, two new area AB networks are created to indicate the two weeks mentioned above by adjusting demands with season trends and adding pump flow based on SCADA data, and some known unfixed leakages are added to the network as well. Then grid searching is conducted for nodes and emitter coefficient. Nodes are searched one by one in area AB network except for reservoir nodes. Emitter coefficients which are used to simulate leakages are searched at appropriate intervals (based on the size of the leakage flow, i.e. bigger leakage flow has bigger search intervals). A difference between the simulated pressure data is calculated in the same way as before for the observed data. Finally, MSE of two differences, the observed values' difference and the simulated values' difference, is used as evaluation function for this problem, and we could find high probability area of leakage with low MSE value.

$$\mathrm{MSE} = \frac{\sum_{i=1}^{Ns}(\sum_{j=1}^{T}(\Delta Po_{i,j} - \Delta Ps_{i,j})^2)}{Ns \cdot T} \qquad (3)$$

Where Ns is the total number of pressure sensors in area AB, T is the total number of simulation time steps, $\Delta Po_{i,j}$ is the difference of two weeks' observed value at pressure sensor i at time j and $\Delta Ps_{i,j}$ is the difference of two weeks' simulated value at pressure sensor i at time j.

In area C, almost every node has an AMR, which facilitates accurate building of an hydraulic model in EPANET. We select one week when the bust had been occurred and compare the observed and the simulated values to determine the location of the leakage. Secondly, we assign the AMR demand data to nodes and calculate the reservoir head pattern according to the SCADA data of tank level. Thirdly, search for nodes and emitter coefficient and find the difference between the observed pressure value and the simulated value. Finally, we use the same way to calculate MSE of the difference and find high probable area of leakage.

Fig.5 shows the result of the leakage occurred at 20th April 2019, 12:00. The shade in grey in the figure represents the size of the MSE. The node with darker color has smaller MSE. Among them, red, orange, yellow, green and blue indicate the five nodes with the smallest MSE, that is, the area with the highest probability of leakage.

Figure 5. Schematic diagram of leakage isolation.

4. Conclusion

This study presents a method for leakage detection and isolation for the battle of L-town DMA. Firstly, data-driven estimations of flow and pressure are proposed to extract leak-induced data from sensor data. Empirical mode decomposition is performed to obtain the time series trend item and estimate flow under normal conditions. Vector auto regressive is employed to extract the spatial correlation of pressure sensors and estimate the time series data of pressure. Secondly, first difference and weekly difference are calculated to reduce the nonstationary of residual series. For abrupt leakages, the duration and size of leakages are identified by 3-sigma SPC method, and the start time and the end time are matched by analysing the leak size and pressure pattern. For the detection of incipient leakages, which usually yield gentle slope and small quantity of water loss, the window-size slope variation method is used in this study. Thirdly, the emitter coefficient of node in EPANET is set to simulate pipe leakage. The idea of control variates is used to find the two weeks when leakage had occurred and not occurred respectively. The difference between the pressure data of two weeks is calculated to represent the impact of the leakage on the pressure of the network. By the above method, 16 leakages are detected and localized in the battle DMA networks in 2019.

REFERENCES

[1] M. Bakker, E. A. Trietsch, J. H. G. Vreeburg, and L. C. Rietveld, "Analysis of historic leakages and leakage detection in water supply areas of different size", Water Science & Technology: Water Supply, vol. 14, no. 6, pp. 1035-1044, 2014.

[2] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences, vol.454, no.1971, pp.903-995, 1998.

[3] Y.P. Wu, S.M. Liu, X. Wu, Y.F. Liu, and Y.S. Guan, "Leakage detection in district metering areas using a data driven clustering algorithm", Water Research, vol. 100, no. 0043-1354, pp. 28-37, 2016.

[4] Y. P. Wu, S. M. Liu, and X. T. Wang, "Distance-based leakage detection using multiple pressure sensors in district metering areas", Journal of Water Resources Planning and Management, vol. 144, no. 11, pp. 06018009, 2018.

[5] X.T. Wang, G.C. Guo, S.M. Liu, Y.P. Wu, X.Y. Xu and K. Smith, " Burst detection in district metering areas Using Deep Learning Method", Journal of Water Resources Planning and Management, vol. 146, no. 6, pp. 04020031, 2020.

A multistage approach to detect and isolate multiple leakages in district metering areas in water distribution systems

Xiaoting Wang[1*], Junyu Li[2], Xipeng Yu[3], Ziqing Ma[4], Yujun Huang[5]
[1,2,3,4,5] Smart Water Research Center, School of Environment, Tsinghua University, Beijing, China
[1] *wang-xt15@mails.tsinghua.edu.cn;*

SUMMARY

Hydraulic accidents or abnormal situations, also known as leakages, cause not only water losses, but also service interruptions and other negative effects. In order to facilitate the rapid response of water utilities and reduce water losses caused by undiscovered leakages, a timely detection and isolation method is required. To solve the battle problem in L-town, this study investigates the potential of the combination use of data-driven method and hydraulic modelling and proposes a multistage approach, which comprises three stages: estimation, identification and localization. Firstly, empirical mode decomposition and vector auto regressive are performed to identify the trend in flow time series and the spatial correlation of pressure values at different sensors. The two data-driven estimations of flow and pressure are used to extract leak-induced values from the monitoring data. Secondly, first difference and weekly difference are calculated to reduce the nonstationary of residual series. The duration and size of leakages are identified by analysing residuals though three-sigma SPC method. Meanwhile, the window-size slope variation method is proposed to detect the start time and leak size of incipient leakages. Thirdly, emitters are used to represent and simulate pipe leakages in EPANET2. Localization algorithm, which uses the idea of control variates to find two weeks when the leakage had occurred and not occurred, is employed to represent the impact of the leakage on the pressure of the network by calculating the difference between the pressure data of two weeks. By identifying the gaps between the differences of the observed values and of the simulated values, areas with a high probability of leakage occurring are found. The proposed method is applied to L-town network. Results show that 16 leakages are detected and localized in 2019, which demonstrates the effectiveness of the solution for leak detection and localization in water distribution systems.