Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2021.27354

227 (2021) 93–103
July

Development of computational algorithms for daily water leak detection in district metered areas based on the principal component analysis

Suwan Park[a,*], Jae-Hong Ha[b]

[a]*Department of Civil and Environmental Engineering, Pusan National University, Busan 46241, Korea, Tel. +82-51-510-2734; Fax: +82-51-513-9596; email: swanpark@pusan.ac.kr*
[b]*Seoyong Engineering, Yongin 16950, Korea*

Received 4 January 2021; Accepted 1 April 2021

ABSTRACT

Techniques for detecting leakage in water pipe networks have been developed worldwide in order to reduce unaccounted-for water quantity and enhance the reliability of the pipe networks. In this paper computational algorithms utilizing principal component analysis (PCA) were developed so that the algorithms can be used in a realistic water pipe network management situation in which the daily flow data of a district metered area (DMA) are needed to be verified for a possible relation with a water leak incident. For the improvement of the algorithms, it was assumed that a manager of a water pipe network uses these algorithms every day to test if yesterday's inflow data to a DMA were an outlier according to the PCA computational algorithm. The flow data used in this study were analyzed to determine the best flow data size for the field use of the developed PCA algorithm. For various flow data sets, which were defined as the smaller sizes of the flow measured in days than the whole data set available, a reference modeling for the PCA was applied to calculate the model outliers by moving the flow data sets day by day. For each DMA the effective outlier detection rates (EODRs) were calculated for the whole range of the defined time windows. The maximum effective outlier detection rate for a DMA was obtained as the maximum of the calculated EODRs. The process and results of the sensitivity analyses of the model parameters were used to suggest guidance on how to determine model parameters for a given flow data.

Keywords: Principal component analysis; District metered area; Water pipe network; Leak detection; Computational algorithm; Flow data

1. Introduction

Water distribution pipe network (WDPN) is one of the essential infrastructures needed to sustain a modern living. Therefore, proper operation and maintenance of the water pipe networks are crucial for the overall health of the consumers in cities and more loosely populated regions such as rural areas alike.

One of the main challenges faced in the operation and maintenance of WDPN in leakage control. In the past decades, various methods have been developed worldwide for detecting leakage in WDPNs in response to the growing concern regarding many problems associated with water loss in the WDPNs. El-Zahab and Zayed [1] provided an extensive overview of leak detection in general including gas pipe leaks in the historical perspective and proposed a new phase (or category) of leak detection, that is, the identification phase, suggesting the leak detection phases as ILLP, identify-localize-locate-pinpoint.

Detecting leaks based on direct observation on the network status such as flow and pressure, which are commonly classified as hardware methods [2] using infrared thermography camera [3,4] ground-penetrating radar [5], and noise loggers [6] is time-consuming and requires a great

* Corresponding author.

1944-3994/1944-3986 © 2021 Desalination Publications. All rights reserved.

amount of fieldwork with extra costs. Other methods are available for leak detection in WDNs such as transient-based methods, which focus on direct detection of negative pressure waves [7] or burst-induced transient signals [8] and methods utilizing well-calibrated hydraulic models [9,10]. However, as Wu and Liu [2] pointed out, studies on transient-based methods and hydraulic model methods were generally conducted using numerical simulations or under heavily controlled laboratory environments, not in real-life WDNs.

Meanwhile, there are computational pipeline monitoring approaches, which are based on software-based systems or data-driven algorithmic monitoring tools that utilize data collected from limited instrumentation such as flow and pressure measuring devices to detect hydraulic anomalies that may indicate pipe leak or commodity release [11].

Data-driven methods are based on the fact that the statistics of the flow variables through the pipeline change when a leak happens and enable researchers to avoid the complexity of WDN's differential dynamic models. As Nowicki et al. [12] pointed out, data-driven methods for leak detection utilizes the measurement data directly to detect a symptom of leakage quickly. This capability of the data-driven method is based on the characteristics of the method itself which can disregard, to a certain extent, the complexity of network modeling and simplify the problem of describing the system in terms of mathematical models to that of extracting numerically important features of the operational data of the system with the help of statistical modeling methodologies. Contrary to the other methods such as hardware-based methods, transient-based methods, and methods utilizing hydraulic models, data-driven methods have been tested under real-life WDN conditions for validation and performance evaluations.

Extensive reviews on the data-driven methods were conducted by Wu and Liu [2]. According to Wu and Liu [2], the data-driven methods can be categorized as classification methods [13,14], which focus on distinguishing bursts from normal data using mainly artificial neural network (ANN) techniques, prediction-classification methods [15–18], which use normal hydraulic data to build prediction models and various data processing techniques such as ANN, Bayesian Inference System, and adaptive forecasting model, and statistical methods [19,20], which generally incorporate statistical process control theory. Each of the methods of leak detection categorized by Wu and Liu [2] possesses issues and limitations, for example, data required for training and testing models and unbalanced data class sizes for the classification methods, effects of uncertainty in historical data on the accuracy of predicted values for the prediction-classification methods, and inappropriate distribution assumptions on data for the statistical methods.

Although Wu and Liu [2] suggested that the prediction-classification methods are considered as a better tool in supporting the decision-making processes involved in leak detection due to its ability to take uncertainty in prediction and classification by incorporating some probabilistic methods, it is considered that the statistical methods still has room to be further developed and applied to real-life data to find its usefulness in assisting leak detection in WDNs. The principal component analysis (PCA) on which this paper is based is one of the statistical data analysis techniques among the data-driven methods.

Many researchers with data-driven methods [21–26] have utilized pressure data of WDNs due to the relatively low costs and easy operability of pressure measuring devices [27]. However, flow is accepted as a more effective parameter for leak detection [28,29].

The abnormal state of WDN are usually caused by device faults (e.g., sensor or pump break down), water leakage in the pipe, a significant increase of the water uptake (e.g., caused by fire brigades), etc. In many situations, analysis of the cause and effects of the abnormal state of WDNs is required to deal with a large amount of data regarding the operational status of the networks that are hard to handle and process. Thanks to the characteristics of the PCA, a multivariate statistical analysis method, that can extract essential information embedded in a large amount of data and indicate the abnormal state of the system using the calculated measures (T^2, SPE and DMOD statistics), it has been used successfully in various applications of pattern recognition and fault diagnosis Gertler [30] and utilized for the processing of the data regarding the operations of WDNs in recent years.

Since the first application of the PCA technique for leak detection by Palau et al. [31], several similar research cases have appeared. Kazimierz et al. [32] used the idea of applying several regional PCA models (PCA monitoring models) for a WDN identified on the basis of spatially local and available measurements to conclude about the operational state of a WDN, instead of a single global model. The main idea of MultiRegional Principal Component Analysis was presented on example of a small water network and the method was applied to DWDS in Chojnice, Northern Poland.

Adam and Michał [33] described an approach to detect leakages in water distribution systems using kernel principal component analysis (KPCA), which can be considered as a non-linear extension of the PCA method and an example of machine learning, with a limited number of measurements. Nowicki et al. [12], based on Adam and Michał [33], presented a systematic and comprehensive approach to use KPCA for fault detection with regard to water leakage and provided a quantitative performance comparison between PCA and KPCA using a hydraulic model of a WDN for a town in Poland which was used to generate values of flows and pressures in monitoring nodes in place of measurement equipment both during correct operation and simulated faults.

Santos-Ruiz et al. [34] proposed a dynamic PCA-based methodology for the detection and quantification of leaks in an experimental pressurized pipeline. The technique was based on an exploratory data analysis of the residuals that result from projecting pressure and flow measurements at the pipeline ends onto the principal and complementary PCA subspaces.

Gertler et al. [35] applied the PCA modeling technique for fault diagnosis in water distribution systems based on the analysis of pressure variations produced by a leakage in the water distribution network. The leakage detection procedure was performed by comparing real pressure and flow data with their estimation using the simulation of the mathematical network model as suggested by Pudar and Liggett [36]. The technique was applied to a simple hydraulic case study and its particularities and the detection results were presented.

Criticizing in the limitations of the works by [12,32,35,37,38] imposed by the assumptions on the statistical properties and the consideration on the uncertainty of the network variables, Quiñones et al. [39] presented evaluation results with regard to the effectiveness of three variants of the PCA technique on the performance of leak detection in WDNs when only a limited number of pressure measurements are available and assuming an uncertain demand that varies over time with white noise in the measurements. The results were obtained from the simulation of an artificial network with similar and different patterns for each demand node.

As described above the PCA technique has been mainly applied to artificial pipe networks or mathematical network simulation models. In this paper, the PCA technique was applied to the recorded inflow data and historical leakage records of district metered areas (DMAs) in South Korea to evaluate its potential in detecting leaks or bursts in the DMAs. Computational algorithms were developed to perform the evaluation by setting up the condition in which realistic application scenarios of the PCA technique was applied. The computational algorithms developed in Park et al. [40] based on the PCA were further modified and enhanced so that the algorithms can be used in a realistic water pipe network management situation in which the daily inflow data of a DMA are needed to be verified for a possible relation with a water leak incident. The process and results of the sensitivity analyses of the model parameters were used to suggest guidance on how to determine model parameters for a given flow data to maximize the leak detection potential of the technique. The developed computational algorithms were coded as self-developed scripts in the MATLAB environment.

2. Material and methods

2.1. Criterion for determining an outlier for a new data set under the PCA modeling

PCA is one of the analytical techniques for analyzing multivariate data. It is a technique for converting multidimensional data to low-dimensional data with minimal loss of information. The principal concept of PCA is to represent the whole information through fewer variables than the original data. Principal components are statistically independent of each other and there is no loss of information when all the main components of the data are used. The first principal component best describes the variability of the data, and the explanatory powers of the principal components diminish gradually.

The PCA consists of a score matrix T of $n \times f$ dimensions and a loading matrix P of $m \times f$ dimensions, where n is the number of observations and m the number of variables. Eq. (1) shows that an original data matrix may be factorized into loading matrix P of $m \times f$ dimensions and a residual matrix E of $n \times m$ dimensions where f is the number of principal components, and $f < m$.

$$X = T \cdot P^T + E \tag{1}$$

The column of the loading matrix P is called a 'principal component' and represents an eigenvector for the eigenvalue of the variance-covariance matrix of X. The eigenvectors are arranged in the order of the corresponding eigenvalues, and the principal components can be selected only partially according to the purpose of analysis. The most optimal partitioning of the raw data is to minimize the residual matrix by partitioning.

Usually, a portion of principal components is selected to convert multidimensional data to low dimensional data. The explanatory power of a PCA model depends on the number of selected principal components for the loading matrix of a finalized model. Cumulative percent variance (CPV) is used to present the explanatory power of a PCA model which is calculated as:

$$\text{CPV} = \frac{\sum_{i=1}^{A} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{2}$$

where λ is the eigenvalue of original matrix X, m is the number of variables, A is the number of principal components which is usually less than m. The number of principal components, A, is usually chosen as 70% or 80% for an analysis.

The Hotelling T^2 statistic or DMOD statistics may be used to determine outliers of a PCA model. According to Palau et al. [31], the Hotelling T^2 statistic is more suitable for detecting abnormal demand in a pipe network, and the DMOD statistic is more suitable for detecting leakage in a pipe network. The equation for obtaining a DMOD statistic is given by Eqs. (3) and (4), respectively.

$$S_i = \sqrt{\frac{\sum_{k=1}^{K} e_{ik}^2}{(K - A)}} \tag{3}$$

$$S_0 = \sqrt{\frac{\sum_{i=1}^{N}\sum_{k=1}^{K} e_{ik}^2}{(N - A - A_0)(K - A)}} \tag{4}$$

where S_i represents the absolute distance to model and S_0 the normalized distance of the model, K the number of primitive variables, A the number of principal components, N the number of observations, e_{ik} residual of i-th observation of variable K, and A_0 1 when it is normalized and 0 otherwise. Eq. (5) represents the criterion for determining whether an estimated DMOD statistic is an outlier.

$$\left(\frac{S_i}{S_0}\right)^2 > F_{\text{critical}} \tag{5}$$

where $(S_i/S_0)^2$ has an F-Snedecor probability distribution with (N-A-1) (K-A) degrees of freedom and (K-A) as parameters. For example, if the calculated DMOD statistic $(S_i/S_0)^2$ is larger than the p-value of the F-Snedecor probability distribution, a DMOD statistic of the flow data is determined to be an outlier. In this study, as shown in Mpesha et al. [8], the

DMOD outlier was used to determine if a set of 24 h inflow data of a DMA represent the occurrence of an anomaly in the DMA of interest.

Once a PCA model is prepared, it can be used to verify whether a new data set belongs to the category of outliers of a constructed PCA model. This verification is first conducted using the loading matrix of a PCA model (P) to calculate the score matrix for a new data set as:

$$T^{\text{new}} = X^{\text{new}} \cdot P \tag{6}$$

where X^{new} is a new data set to be verified. The residual matrix for a new data set is calculated using Eq. (7).

$$E^{\text{new}} = X^{\text{new}} - T^{\text{new}} \cdot P^T \tag{7}$$

The DMOD statistics for the new data set is calculated as Eq. (8).

$$\text{DMOD}^{\text{new}} = \left(\frac{S_i^{\text{new}}}{S_0} \right)^2 \tag{8}$$

The new data set is determined to be an outlier of a constructed PCA model if the DMOD statistics for the new data set satisfies the following:

$$\text{DMOD}^{\text{new}} > F_{\text{critical}} \tag{9}$$

2.2. Flow data and historical leakage records

This study used the inflow measurement data and records of historical leak incidents in a water pipe network in South Korea. The flow data utilized in this study were the 24 h flow data of 26 DMAs recorded every hour for 843 d, which were from October 28, 2016 to February 17, 2019. The historical leakage records consisted of the dates and locations of the leak incidents in the network. The leakage records contained 120 dates which represented the dates when leakage repair was completed. Table 1 shows a portion of the DMA inflow data of the water pipe network.

2.3. Developed computational algorithms for leak detection

In this study, the computational algorithms developed in Park et al. [40] were modified to verify and enhance the utility of the algorithms in detecting leak events in the real-life maintenance situation of a WDPN. Park et al. [40] used the 426 d of the hourly inflow data of 11 DMAs to analyze any relevance between the recorded leak events and the calculated outliers of the constructed PCA model for the flow data in the WDPN under study. In Park et al. [40] the potential of the PCA in detecting leaks in a WDPN was verified by analyzing the values of the computed maximum effective outlier detection rate (M-EODR) and the variability of the best time window (BTW) for which the M-EODR for each DMA understudy was calculated.

However, the method used in Park et al. [40] to compute the M-EODRs and BTWs did not consider the actual circumstance inherent in the managerial decision-making processes in the management of WDPNs. That is, the managers may need to make a decision regarding the status of the flow data on daily basis to prevent a leak or minimize the severity of a possible leak event. In other words, the managers of WDPNs may be confronted with making a proper decision whether the current flow data is abnormal and a corresponding leak detection work order needs to be issued.

In this regard, the computational algorithms developed in Park et al. [40] were modified in this study to verify and enhance the applicability of the PCA assuming the managers in the field run the algorithms every day to check if the latest flow data is an outlier of a constructed PCA model. For this verification and enhancement of the algorithms, the flow data used in this study were analyzed to determine the best flow data size to run the algorithm for the field use of the developed PCA algorithms. To determine the best flow data size various flow data sets (FDSs) have been tried. A FDS was designed to be identified as the size. For example, a FDS may be consisted of 100 d or another of 110 d of flow data. A sub-category of FDS was designed so that a FDS could occupy a certain range of time for the reference PCA modeling and test for outliers. For example, a FDS of 100 d could contain flow data from June 1, 2017 to August 8, 2017 or from June 2, 2017 to August 9, 2017.

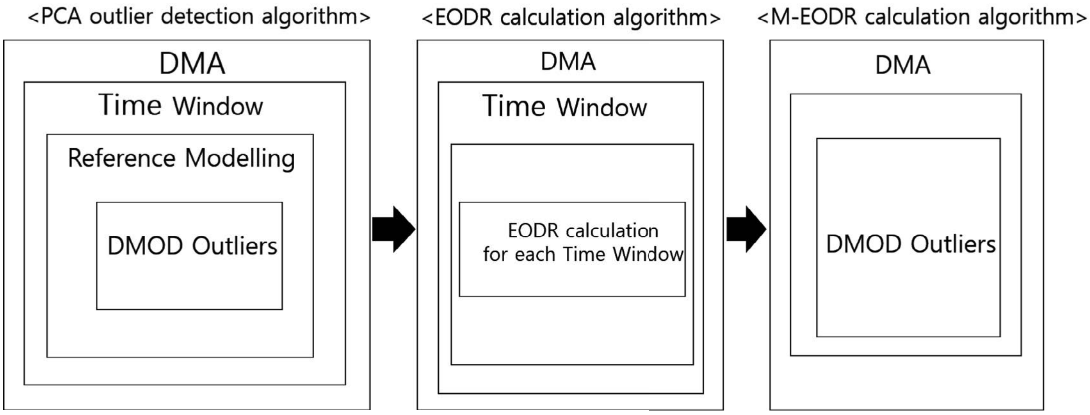
The sizes of FDS were defined to be smaller, which was measured in days than the whole data set available. The outliers in a FDS were removed to establish a reference PCA model for a FDS. Once a reference model of a FDS was constructed, the flow data in the next 24 h following the FDS under consideration, which was termed as 'flow data to be verified (FDV)', was examined using the PCA technique to determine the FDV of interest was an outlier of the reference PCA model of the FDS. After the test for outlier for a FDV was finished, a FDS of a particular size was moved by 1 d while maintaining its size until there were no FDVs left in the available flow data used for the analysis.

The effective outlier detection rate (EODR) defined in Park et al. [40] and shown as Eq. (10) was calculated based on the calculated outliers of a FDS and the historical leak records of a DMA.

$$\text{EODR}(\%) = \left(\frac{\text{number of outliers in ODP}}{\text{total number of outliers}} \right) \times 100 \tag{10}$$

ODP in Eq. (10) means 'outlier detection period' that was used to distinguish valid or effective outliers in calculating EODRs. An outlier was considered as an effective outlier if it is inside an ODP which was defined to be a finite number of days, for example, 15 d, and in which the date of completion of leakage repair was centered. 'Total number of outliers' and 'number of outliers in ODP' were calculated after testing all of the FDVs for the occurrence of outliers in the whole flow data available for a DMA by advancing a FDS by one day along the timeline after each test for a FDV.

For each DMA the EODRs were calculated for the whole range of the defined time windows described in Park et al. [40]. The M-EODR for a DMA was obtained as the maximum of the calculated EODRs. Fig. 1 shows a block representation of the algorithms for calculating the M-EODR for the field use of the PCA. Each block in Fig. 1 represents the sub-algorithm for repeatedly calculating the item in a block.



<PCA outlier detection algorithm>
DMA
Time Window
Reference Modelling
DMOD Outliers
<EODR calculation algorithm>
DMA
Time Window
EODR calculation for each Time Window
<M-EODR calculation algorithm>
DMA
DMOD Outliers

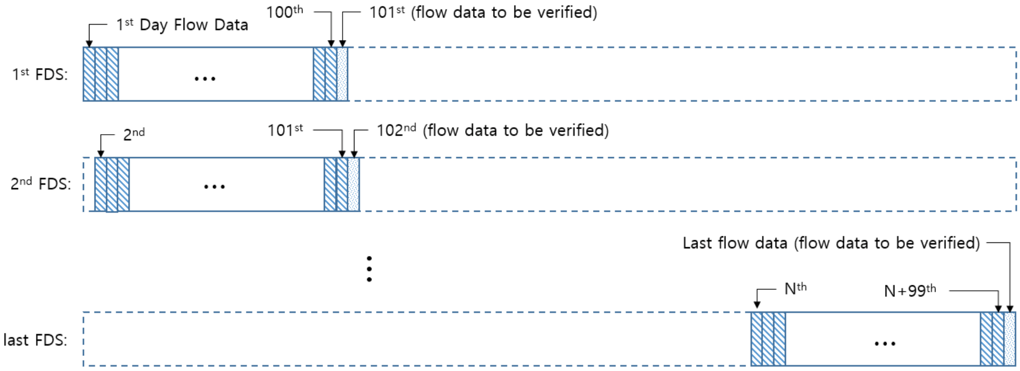
Fig. 1. Block representation of the developed algorithms.

Table 1
Sample inflow data in the case study area WDPN

Date	Hour	Inflow to DMA-1 (m^3/h)	Inflow to DMA-2 (m^3/h)	Inflow to DMA-3 (m^3/h)
2017-06-19	23:00	236	99	138
2017-06-19	24:00	170	83	133
2017-06-20	1:00	130	71	113
2017-06-20	2:00	140	59	17
2017-06-20	3:00	192	54	49
2017-06-20	4:00	378	50	15
2017-06-20	5:00	186	40	17
2017-06-20	6:00	346	53	242
2017-06-20	7:00	710	81	258
2017-06-20	8:00	516	97	293
2017-06-20	9:00	276	92	357

As shown in Fig. 2, a FDS of 100 d (FDS_100) was used to illustrate the process of the PCA modeling and testing of outliers for the FDVs of a DMA. The first FDS of the FDS_100 was prepared by excluding the whole flow data in a day if a portion of the flow data in a day is missing or has abnormal values. Then, the first FDS was used to build a reference PCA model and the flow data of the next day, which is the first FDV (i.e., flow data in the 101st day in Fig. 2), were checked to find out if they were an outlier of the reference PCA model for the first FDS. Then, the first FDS moves to the next day to form the second FDS of the FDS_100 while retaining the same number of days, which is 100 d, as the first FDS. Therefore, the second FDS consisted of the flow data from the second day until the 101st day of available flow data. A reference PCA modeling technique was applied to the second FDS to find out if the flow data in the second FDV, which is the 102nd day, were an outlier of the reference PCA model constructed using the second FDS. This process continues until the FDV reaches the end of the whole flow data available.

The algorithms were designed so that in case a FDV contains missing values they were treated as outliers of the reference PCA models of the FDS's. These outliers were assigned a label and were not counted as valid outliers, thus, were not used in the calculation of the EODRs. These processes of progressing the FDS and calculating the EODRs of a DMA were conducted for all of the time windows designed for the analysis. The reference PCA model of a FDS was built by applying the PCA algorithm to a FDS repeatedly until all outliers in a FDS were removed from a FDS. Fig. 3 shows the process of the reference PCA modeling utilized in this study.



1st Day Flow Data
100th
101st (flow data to be verified)
1st FDS:
2nd
101st
102nd (flow data to be verified)
2nd FDS:
Last flow data (flow data to be verified)
Nth
N+99th
last FDS:

Fig. 2. Progression process of a FDS for the PCA outlier detection and EODR calculation.

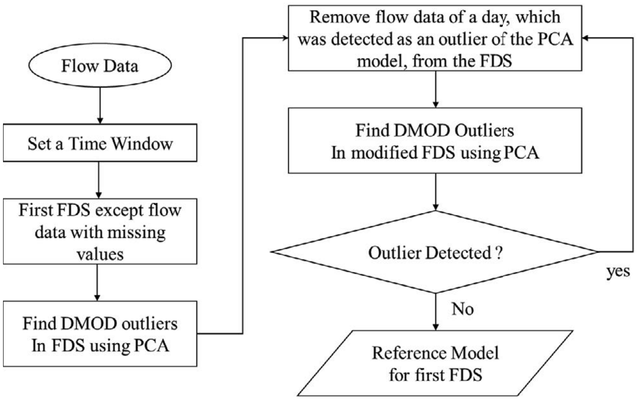

Flow Data
Set a Time Window
First FDS except flow data with missing values
Find DMOD outliers In FDS using PCA
Remove flow data of a day, which was detected as an outlier of the PCA model, from the FDS
Find DMOD Outliers In modified FDS using PCA
Outlier Detected ?
yes
No
Reference Model for first FDS

Fig. 3. Process of a PCA reference modeling for the first FDS.

The algorithm in Fig. 4 is continued from Fig. 3 and describes the procedure of advancing a FDS and FDV by 1 d until the FDV reaches the end of available flow data. For each FDS in Fig. 2, the algorithm shown in Fig. 4 was applied to determine whether a FDV is an outlier based on the reference PCA model of the corresponding FDS.

The algorithms in Figs. 3 and 4 were conducted for all of the time windows defined using the method of Park et al. [40]. The time windows were defined using 'center time' and 'time range' to extract a portion of the hourly flow data. The time range was increased from 3 to 23 h in the increment of 2 h. The center time of each time window was varied from 1 to 24 o'clock. Fig. 5 shows three examples of the time windows that can be defined for the analyses in this study.

3. Results and discussion

3.1. Sensitivity analyses of the model parameters

The process and results of the sensitivity analyses of the model parameters were used to suggest guidance on how to determine model parameters and, consequently, the best flow data size (the final size of a FDS) to obtain the best results of the analyses for a given flow data and historical leak records. It was considered that the manager of a WDPN will want to use the most current flow data for calculating the M-EODR of a DMA due to a fluctuating water demand trend in the DMA of interest. In this case, the size of a FDS plays an important role in the calculation of M-EODR. Therefore, the sensitivity of the size of FDS on M-EODR was first conducted among the parameters of the reference PCA model, which are the CPV, p-value of the F-distribution which was used for the calculation of the critical DMOD, and the size of FDS.

For fixed values of the CPV and p-value of the F-distribution analyses of the changes in the M-EODRs of a DMA for various sizes of the FDS were conducted to determine the best FDS size of a DMA. In addition, the manager may need to have his/her own criterion regarding the value of the M-EODR to decide whether a preemptive leak detection work order needs to be issued. In this case, the manager may choose the minimum value of the M-EODR to issue the work order as close to 100% if he/she wants to save the budget for preemptive maintenance as much as possible. In this study, the minimum value of the M-EODR to issue the work order was considered as 60%. Based on this rule three M-EODR categories were defined to analyze the effects of the changes in the parameters. The categories were set as M-EODR equal to 100%, M-EODR between 60% and 100%, M-EODR less than 60%.

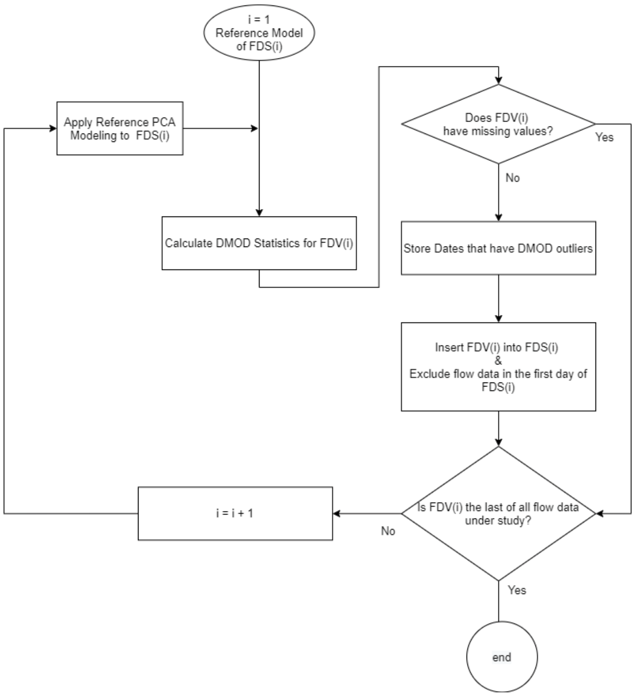

i = 1
Reference Model of FDS(i)
Apply Reference PCA Modeling to FDS(i)
Does FDV(i) have missing values?
Yes
No
Calculate DMOD Statistics for FDV(i)
Store Dates that have DMOD outliers
Insert FDV(i) into FDS(i) & Exclude flow data in the first day of FDS(i)
i = i + 1
Is FDV(i) the last of all flow data under study?
No
Yes
end

Fig. 4. Computational algorithms for calculating outliers of the reference PCA models for the FDS's of a DMA.

The numbers of DMAs that belong to the M-EODR categories in Fig. 6 were calculated using the reference modeling technique shown in Figs. 3 and 4 for the FDS sizes of 30, 60, 90, 120, and 150 d. For example, for the FDS size of 150 d, 10 DMAs were found to have the M-EODR of 100%.

Although the numbers of DMAs belonging to each M-EODR category in Fig. 6 are similar for the various FDS sizes used, the manager of a WDPN was assumed to take the most appropriate FDS size for further analyses. It was considered rational that the best FDS size was the one resulting in the greatest number of DMAs with the calculated M-EODRs greater than a preset minimum value of the M-EODR of 60%. Therefore, a FDS of 90 d of flow data was chosen for the flow data used in this study. The sensitivities of the CPV and the p-value of the F-distribution on the M-EODRs were analyzed for the case of 90 d flow data.

3.2. Analysis of the modeling results

Table 2 shows the M-EODRs, the number of leak incidents and a corresponding number of effective outliers obtained using the flow data and computational algorithms developed in this study with the ODP of 15 d. The default values of the parameters that are the FDS size of 90 d, CPV of 70% and p-value of 0.05 were used in the computation.

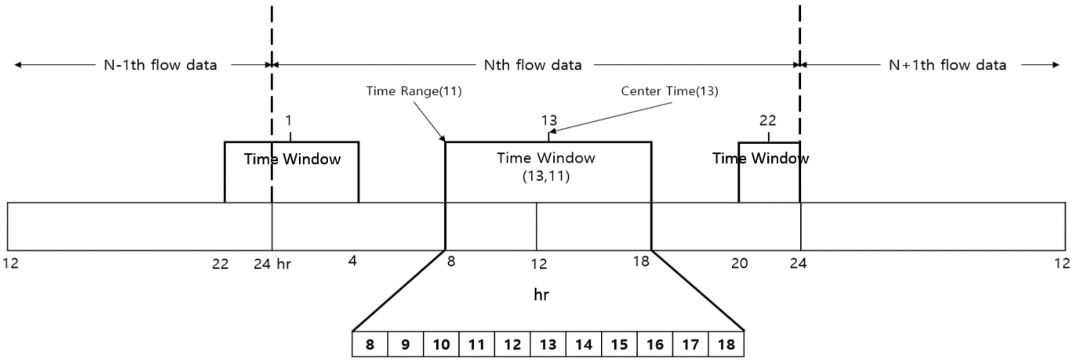


Fig. 5. Example of three-time windows.

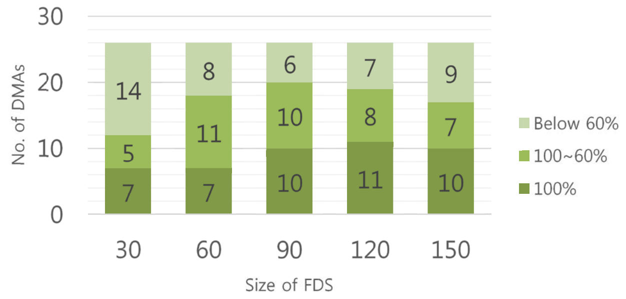


Fig. 6. Numbers of DMAs that belong to the M-EODR categories for various FDS sizes for CPV of 70% and *p*-value of 0.05.

As shown in Table 2, the case of M-EODR equal to 100% usually resulted in a very low number of outliers compared to the number of leak incidents. For example, there was a DMA with 66 leak incidents and two outliers of the reference PCA model which are all inside the defined ODPs. Therefore, although the performance of the developed algorithm was turned out to be excellent in detecting leak events for this DMA, the reliability of the calculated M-EODR of 100% may need to be further investigated and it may be decreased if further new FDS's are collected in the future and used for the analysis. Therefore, it was considered that the highest EODR next to 100% should be chosen for determining managerial decision whether a fieldwork order for leak detection needs to be issued. For example, the manager of a water pipe network may decide not to issue a field leak detection work order for a DMA if the highest value of the EODR of a DMA excluding 100% is less than 60% even if the flow data of a previous day of the DMA of interest is calculated as an outlier.

Table 3 shows the results of the analysis similar to Table 2 except that Table 3 shows the highest EODR next to 100% as the M-EODR of a DMA using the default model parameter values. The highest EODR next to 100% is termed as 'practical maximum effective outlier detection rate (P-M-EODR)' in the following description. In Table 3, the values of the center time and time range represent the time windows of a DMA for the corresponding P-M-EODR.

As shown in Fig. 7, the number of DMAs with the M-EODR of greater than 60% changes in great number from the CPV of 50% to 90%. Although the number of DMAs with the M-EODR of greater than 60% was the greatest for the CPV of 90%, the number of DMAs with the M-EODR of 100% was relatively high for the CPV of greater than or equal to 90%. Moreover, for the CPV of greater than or equal to 90% the PCA reference modeling encountered a problem of no DMOD statistics calculated due to the number of the principal components being equal to the number of variables. Additionally, for the case of the CPV of greater than or equal to 90%, the M-EODR became either 0 or 100% for each Time Window. Therefore, due to the aforementioned problem of the M-EODR of 100% having a very little number of effective outliers compared to the number of leak incidents and the numerical problem of calculating the DMOD statistics for the CPV of over 90%, it was considered that the CPV of 70% should be chosen as a default value of the CPV in this study for more effective applications of the developed technique.

Table 2
M-EODRs for the DMAs

DMA	1	2	3	4	5	6	7	8	9	10	11	12	13
M-EODR (%)	55	100	100	60	100	100	100	100	100	100	50	86	51
No. of leak records	38	66	41	12	7	36	70	22	60	41	11	10	32
No. of effective outliers	28	2	1	9	1	1	1	1	1	2	6	6	104
DMA	14	15	16	17	18	19	20	21	22	23	24	25	26
M-EODR (%)	75	77	100	67	61	62	64	60	32	100	64	33	49
No. of leak records	17	41	53	14	22	23	29	23	12	30	18	8	19
No. of effective outliers	3	10	11	2	17	34	14	3	46	1	49	1	49

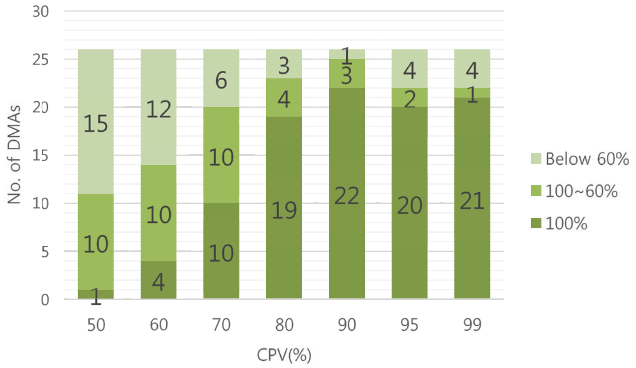


Fig. 7. Changes of the M-EODRs according to different values for the CPV using the p-value of 0.05.

As shown in Fig. 8 the number of DMAs with the M-EODR of greater than 60% was increased from the case of the p-value of 0.07 to 0.01. Although the number of DMAs with the M-EODR of greater than 60% was the greatest for the p-value of 0.01, the number of DMAs with the M-EODR of 100%, in that case, was relatively high compared to other cases of the p-values. For the cases of the p-value of 0.05 and 0.03, the number of DMAs that have the M-EODR of greater than or equal to 60% was the same. Therefore, the p-value of 0.05 may be chosen by a manager of a water distribution system for a more effective application of the technique since the number of DMAs with the M-EODR of 100% is smaller than that of the p-value of 0.03.

Figs. 9 and 10 show the time windows and center times of the DMAs with the P-M-EODR greater than 60% in Table 3. Since the time windows shown in Figs. 9 and 10 correspond to the analysis time interval that has the maximum number of effective outliers in a day within an ODP, the time windows that has the P-M-EODR are considered to be representing the approximate time zone in a day in which abnormal flows related to leaking incidents usually occurred for the DMA of interest. For this case study, the center times of about 63% of the DMAs were found to be between 9 and 12 o'clock. However, the exact timing of the occurrences of the leak events and the occurrences of the outliers need to be checked and analyzed using leak records with more detailed leak-related information to see if there are any delays in the occurrences of the outliers after the occurrence of the leak events.

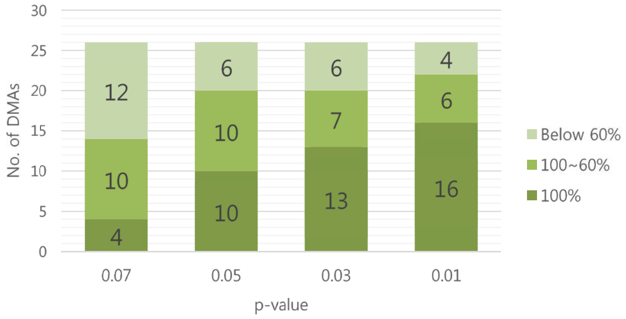


Fig. 8. Changes of the M-EODRs according to different values for the p-value using the CPV of 70%.

Due to the characteristics of the PCA, it was found out that two one-day flow data may have different outlier detection results. In other words, water flow data on a specific day may be calculated as an outlier of the reference PCA modeling analysis and another flow data on a different day with a similar pattern and amount may not turn out to be an outlier of the analysis. Fig. 11 shows an example of the two similar flow data, that is the 258th day (July 07, 2017) and the 265th day (July 19, 2017) flow data, which have different results of the outlier detection.

After applying the outlier calculation algorithm, the 258th day (July 7, 2017) flow data turned out to be an outlier while the 265th day (July 19, 2017) flow data turned out not to be an outlier with regard to the critical DMOD statistics value for outlier detection, which is 1.61 for both days.

This phenomenon of the outlier calculation results is due to the model error calculation process shown in Eqs. (6)–(8). Since the loading matrix of a FDS reflects the low dimensional characteristics of the whole FDS and changes depending on where it is located in time, the residual error matrix of a FDV, which is E^{new} in Eq. (7), may be different even for the same total amount and pattern of water flow data. This leads to the changes in the corresponding value

Table 3
P-M-EODRs for the DMAs

DMA	1	2	3	4	5	6	7	8	9	10	11	12	13
P-M-EODR (%)	55	95	96	60	67	91	92	86	94	85	50	86	51
Center time	11	11	10	10	12	1	12	11	21	12	15	10	17
Time range	3	7	17	7	5	19	3	23	19	9	3	5	21
No. leak records	38	66	41	12	7	36	70	22	60	41	11	10	32
No. of effective outliers	28	42	22	9	4	10	11	6	16	11	6	6	104
DMA	14	15	16	17	18	19	20	21	22	23	24	25	26
P-M-EODR (%)	75	77	94	67	61	62	64	60	32	57	64	33	49
Center time	24	10	19	9	5	23	9	10	11	20	6	9	18
Time range	3	5	3	3	3	3	5	7	7	3	3	3	23
No. leak records	17	41	53	14	22	23	29	23	12	30	18	8	19
No. of effective outliers	3	10	16	2	17	34	14	3	46	17	49	1	49

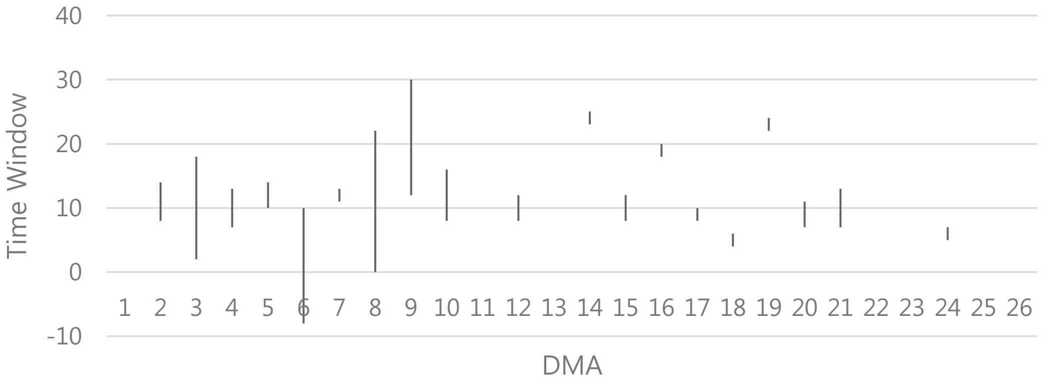


Fig. 9. Time windows resulted in the P-M-EODR of each DMA.

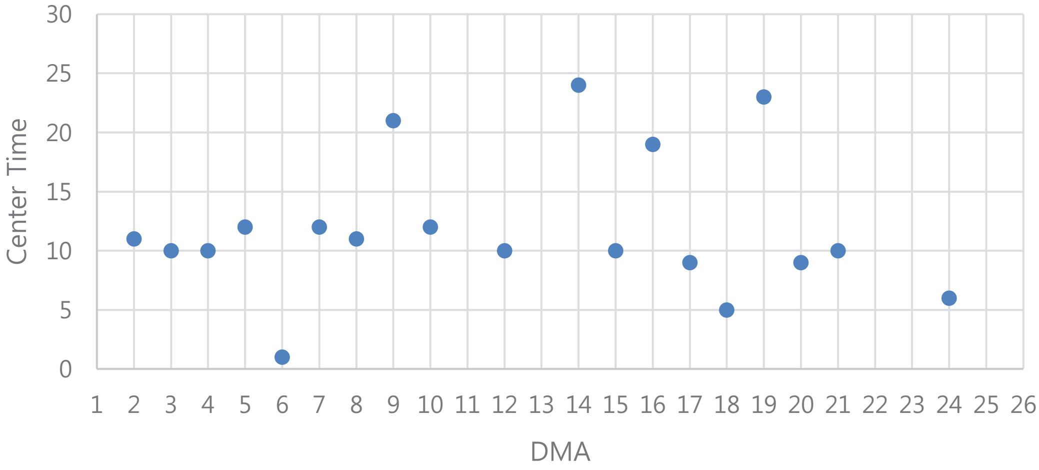


Fig. 10. Center times of the time windows resulted in the P-M-EODR of each DMA.

of the DMOD statistics of a FDV. Therefore, it is likely that a similar daily flow amount and pattern may or may not be detected as an outlier. This characteristic of the PCA modeling technique is considered to be useful in detecting an anomaly in the flow data of a DMA since it uses the variations in the flow as a whole to detect a new anomaly of flow.

4. Conclusions

In this study, the computational algorithms developed in Park et al. [40] based on the PCA were further modified and enhanced so that the algorithms can be used in a realistic water pipe network management situation in which the daily inflow data of a DMA are needed to be verified for a possible relation with a water leak incident. For the improvement of the algorithms, it was assumed that a manager of a water pipe network will use these algorithms every day to test if yesterday's inflow data to a DMA were an outlier according to the PCA computational algorithm which would mean that the statistically abnormal flow data of a previous day may be a result of a possible leak event in the network.

Three M-EODR categories were defined to analyze the effects of the changes in the parameters. The categories were set as M-EODR equal to 100%, M-EODR between 60% and 100%, M-EODR less than 60% based on the assumption that the minimum value of the M-EODR to issue the work order may be set as 60%. The process and results of the sensitivity analyses of the model parameters were used to suggest guidance on how to determine model parameters for a given flow data. A rule of thumb that can be used for selecting appropriate values of the parameters was determined

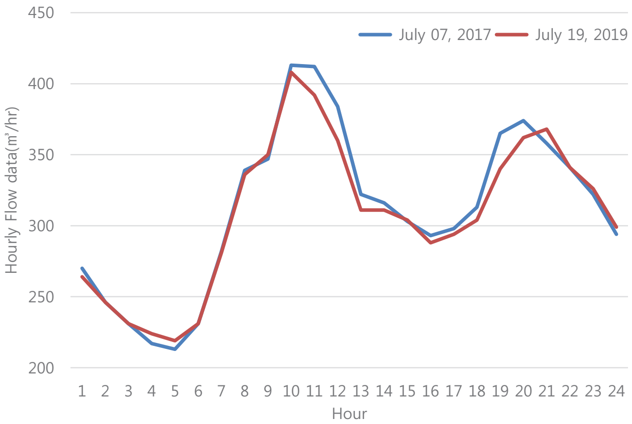


Fig. 11. Example of the two similar flow data patterns which have different results of the outlier detection.

to select a parameter value that maximizes the number of DMAs with M-EODR between 60% and 100%.

For the flow data used in this study, a FDS with 90 d of flow data was chosen for the most effective leak detection analysis of the case study flow data. It was considered that the highest EODR next to 100% should be chosen for determining managerial decision whether a fieldwork order for leak detection needs to be issued due to the reliability of the calculated M-EODRs of 100%. In this study, the minimum value of the M-EODR to issue the work order was considered as 60%.

The analysis results showed that the range of the M-EODR except 100% was from 32% to 96% among which 73% of the DMAs had the values of M-EODR as over 60%. Therefore, it was concluded that for the DMAs with the values of M-EODR as over 60% the computational algorithms developed in this study may be used to assist the manager of a case study network in deciding whether a further leak detection field-work order needs to be issued based on the calculation results of the developed algorithms. For DMAs with higher M-EODR may give more confidence to the manager in deciding to send field crews for finding the leak and its location. Analysis of the characteristics of the PCA that produce different outlier calculation results due to the model error calculation processes for the new data set was also provided.

Since the analysis results greatly depend on the accuracy of the data used, DMAs with more accurate leak records are expected to produce even better results. The developed algorithm may be modified further to analyze flow data recorded every 1 min or less instead of an hour to assist in the managerial decision-making processes to detect and cope with more serious pipe burst accidents. A deeper connection between the timing of the occurrences of the leak events and the occurrences of the outliers such as delay in the occurrences of the outliers after the occurrence of leak events may be revealed in future studies utilizing more detailed leak-related information.

Acknowledgments

This subject is supported by the Korea Ministry of Environment as “Global Top Project (code 2016002130005)”.

References

[1] S. El-Zahab, T. Zayed, Leak detection in water distribution networks: an introductory overview, Smart Water, 4 (2019), doi: 10.1186/s40713-019-0017-x.

[2] Y.P. Wu, S.M. Liu, A review of data-driven approaches for burst detection in water distribution systems, Urban Water J., 14 (2017) 972–983.

[3] M. Fahmy, O. Moselhi, Automated detection and location of leaks in water mains using infrared photography, J. Perform. Constr. Facil., 24 (2010), doi: 10.1061/(ASCE) CF.1943-5509.0000094.

[4] A. Al Hawari, M. Khader, T. Zayed, O. Moselhi, Non-destructive visual-statistical approach to detect leaks in water mains, World Acad. Sci. Eng. Technol., Int. J. Environ. Ecol. Eng., 9 (2015) 230–234.

[5] A. Al Hawari, M. Khader, T. Zayed, O. Moselhi, Detection of leaks in water mains using ground penetrating radar, World Acad. Sci. Eng. Technol., Int. J. Environ. Ecol. Eng., 10 (2016) 422–425.

[6] S. El Zahab, F. Mosleh, T. Zayed, An accelerometer-based real-time monitoring and leak detection system for pressurized water pipelines, Pipelines, 2016 (2016) 257–268.

[7] S.J. Lee, G.B. Lee, J.C. Suh, J.M. Lee, Online burst detection and location of water distribution systems and its practical applications, J. Water Resour. Plann. Manage., 142 (2016), doi: 10.1061/(ASCE)WR.1943–5452.0000545.

[8] W. Mpesha, S.L. Gassman, M.H. Chaudhry, Leak detection in pipes by frequency response method, J. Hydraul. Eng., 127 (2001) 55–62.

[9] R. Pérez, G. Sanz, V. Puig, J. Quevedo, M.À. Cugueró-Escofet, F. Nejjari, J. Meseguer, G. Cembrano, J.M. Mirats-Tur, R. Sarrate, Leak localization in water networks: A model-based methodology using pressure sensors applied to a real network in barcelona [applications of control], IEEE Control Syst. Mag., 34 (2014) 24–36.

[10] G. Sanz, R. Pérez, Z. Kapelan, D. Savic, Leak detection and localization through demand components calibration, J. Water Resour. Plann. Manage., 142 (2016), doi: 10.1061/(ASCE) WR.1943-5452.0000592.

[11] API, Computational Pipeline Monitoring for Liquid Pipelines, American Petroleum Institute (API), 2002.

[12] A. Nowicki, M. Grochowski, K. Duzinkiewicz, Data-driven models for fault detection using kernel PCA: a water distribution system case study, Int. J. Appl. Math. Comput. Sci., 22 (2012) 939–949.

[13] K. Aksela, M. Aksela, R. Vahala, Leakage detection in a real distribution network using a SOM, Urban Water J., 6 (2009) 279–289.

[14] S.R. Mounce, J. Machell, Burst detection using hydraulic data from water distribution systems with artificial neural networks, Urban Water J., 3 (2006) 21–31.

[15] C.J. Hutton, Z. Kapelan, Real-time burst detection in water distribution systems using a Bayesian demand forecasting methodology, Procedia Eng., 119 (2015) 13–18.

[16] C.J. Hutton, Z. Kapelan, A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting, Environ. Modell. Software, 66 (2015) 87–97.

[17] M. Romano, K. Woodward, Z. Kapelan, D.A. Savić, Near Real-time Detection of Pipe Burst Events in Cascading District Metered Areas, 11th International Conference on Hydroinformatics, HIC 2014, New York City, USA, 2014.

[18] M. Bakker, J.H.G. Vreeburg, K.M. van Schagen, L.C. Rietveld, A fully adaptive forecasting model for short-term drinking water demand, Environ. Modell. Software, 48 (2014) 141–151.

[19] D.H. Jung, K. Lansey, Water distribution system burst detection using a nonlinear kalman filter, J. Water Resour. Plann. Manage., 141 (2015), doi: 10.1061/(ASCE)WR.1943–5452.0000464.

[20] D. Loureiro, C. Amado, A. Martins, D. Vitorino, A. Mamade, S.T. Coelho, Water distribution systems flow monitoring and anomalous event detection: a practical approach, Urban Water J., 13 (2016) 242–252.

[21] R. Pérez, V. Puig, J. Pascual, A. Peralta, E. Landeros, Ll. Jordanas, Pressure sensor distribution for leak detection in Barcelona water distribution network, Water Sci. Technol., 9 (2009) 715–721.

[22] R. Pérez, J. Quevedo, V. Puig, F. Nejjari, M.A. Cugueró, G. Sanz, J.M. Mirats, Leakage Isolation in Water Distribution Networks: a Comparative Study of Two Methodologies on a Real Case Study, 19th Mediterranean Conference on Control and Automation (MED), Corfu, Greece, 2011, pp. 138–143.

[23] M.V.C. Ponce, L.E.G. Castañón, V.P. Cayuela, Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities, J. Hydroinf., 16 (2014) 649–670.

[24] Y.S. Kim, S.J. Lee, T.Y. Park, G.B. Lee, J.C. Suh, J.M. Lee, Robust leak detection and its localization using interval estimation for water distribution network, Comput. Chem. Eng., 92 (2016) 1–17.

[25] C.C. Sun, B. Parellada, V. Puig, G. Cembrano, Leak localization in water distribution networks using pressure and data-driven classifier approach, Water, 12 (2020) 54, doi: 10.3390/w12010054.

[26] A. Soldevila, J. Blesa, R.M. Fernandez-Canti, S. Tornil-Sin, V. Puig, Data-driven approach for leak localization in water distribution networks using pressure sensors and spatial interpolation, Water, 11 (2019) 1500, doi: 10.3390/w11071500.

[27] P. Ostapkowicz, Leak detection in liquid transmission pipelines using simplified pressure analysis techniques employing a minimum of standard and non-standard measuring devices, Eng. Struct., 113 (2016) 194–205.

[28] S.R. Mounce, R.B. Mounce, J.B. Boxall, Novelty detection for time series data analysis in water distribution systems using support vector machines, J. Hydroinf., 13 (2011) 672–686.

[29] G.L. Ye, R.A. Fenner, Kalman filtering of hydraulic measurements for burst detection in water distribution systems, J. Pipeline Syst. Eng. Pract., 2 (2011) 14–22.

[30] J. Gertler, Fault Detection and Diagnosis, Springer, J. Baillieul, T. Samad, Eds., Encyclopedia of Systems and Control, Springer, London, 2015. Available at: https://doi.org/10.1007/978-1-4471-5058-9_223.

[31] C.V. Palau, F. Arregui, A. Ferrer, Using multivariate principal component analysis of injected water flows to detect anomalous behaviors in a water supply system – a case study, Water Supply IWA, 4 (2004) 169–182.

[32] D. Kazimierz, B. Adam, M. Krzysztof, G. Michal, B.A. Mietek, J. Krzysztof, Leakage detection and localisation in drinking water distribution networks by MultiRegional PCA, Stud. Inf. Control, 17 (2008) 135–152.

[33] N. Adam, G. Michał, Kernel PCA in Application to Leakage Detection in Drinking Water Distribution System, International Conference on Computational Collective Intelligence, ICCCI 2011: Computational Collective Intelligence. Technologies and Applications, Vol. 6922, 2011, pp. 497–506.

[34] I. Santos-Ruiz, F.R. López-Estrada, V. Puig, E.J. Pérez-Pérez, J.D. Mina-Antonio, G. Valencia-Palomo, Diagnosis of fluid leaks in pipelines using dynamic PCA, IFAC-PapersOnLine, 51 (2018) 373–380.

[35] J. Gertler, J. Romera, V. Puig, J. Quevedo, Leak Detection and Isolation in Water Distribution Networks Using Principal Component Analysis and Structured Residuals, 2010 Conference on Control and Fault-Tolerant Systems (SysTol), Nice, France, 2010, pp. 191–196.

[36] R.S. Pudar, J.A. Liggett, Leaks in pipe networks, J. Hydraul. Eng., 118 (1992) 1031–1046.

[37] L. Ferrandez-Gamot, P. Busson, J. Blesa, S. Tornil-Sin, V. Puig, E. Duviella, A. Soldevila, Leak localization in water distribution networks using pressure residuals and classifiers, IFAC-PapersOnLine, 48 (2015) 220–225.

[38] P. Cugueró-Escofet, J. Blesa, R. Pérez, M.A. Cuguero-Escofet, G. Sanz, Assessment of a leak localization algorithm in water networks under demand uncertainty, IFAC-PapersOnLine, 48 (2015) 226–231.

[39] M. Quiñones, C. Verde, A. Prieto, O. Llanes, Detección de fugas enredes de distribución empleando análisis de components principals, Memorias del Congreso Nacional de Control Automático, Querétaro, México, Septiembre 28–30, 2016, pp. 276–282.

[40] S.W. Park, J.H. Ha, K.M. Kim, The principal component analysis for detecting leaks in water pipe networks utilizing flow and leak record data, Desal. Water Treat., 143 (2019) 69–77.