# Leakage Detection and Isolation in Water Distribution Network Based on Data Mining and Genetic Optimized Hydraulic Simulation

**Runze LIU[1], Zhen ZHANG[2], Dian ZHANG[3]**
[1,2,3] DHI China, 4th Floor, Building A, 181 Guyi Road, Shanghai 200235
*[1]lirz@dhigroup.com, [2]zhzh@dhigroup.com, [3]zhdi@dhigroup.com*

## EXTENDED ABSTRACT

Water loss in distribution systems has been a main issue, not only is a waste of water resource and money, but it is also a potential health risk since pollutants could get into pipes through cracks. Leakage detection and management has thus become a critical work for utility companies to reduce water loss. One common practice is based on district metering area (DMA) and night minimum flow (NMF) theories. The NMF is considered stable for a certain part of network, so changes of NMF indicates occurrences of potential leakage in this DMA. Concerned workers could then isolate the exact leaking location by acoustic equipment or get precise geographic information from acoustic monitors installed in network. As the digitalization of urban infrastructures proceeds, SCADA system has been widely used in urban water distribution network management, providing numerous data for analysis. Data-driven models for leakage detection have been proposed since then. Mounce et al (2010) proposed an artificial neural network combined with a fuzzy inference system to detect pipe bursts. Ye & Fenner (2011) proposed an adaptive Kalman filtering method which is more efficient. Other methods such as Bayesian inference system and Golden Section method are also proposed by researchers and engineers. Inspired by such data-driven methods, a new approach for leakage detection is proposed in this abstract, applying data mining models and numerical hydraulic simulations. Data mining models are trained by pressure data, which contains measured SCADA data and hydraulic simulated data. Trained data mining model could detect a certain area where potential leakage exists. Hydraulic simulations are then run by EPANET to simulate leakage events. Genetic algorithm is used to optimize the loss function between simulated results and measured data to find the leakage location and flowrate which matches most. The proposed approach is applied to a hypothesis WDN "L-Town" to detect and isolate the leakage events in 2018 and 2019. The results show that leakage events of 2018 revealed by the proposed approach are highly matched with the actual leakage report provided by "L-Town" with respect to location and time. The detected leakage events of 2019 will be compared with the actual events once the 2019 leakage report is accessible.

### 1. Methodology

#### 1.1 Hydraulic Model Calibration

In order to represent the real-world water distribution network, the mathematical hydraulic model should be calibrated. Typically, parameters to be calibrated include pipe roughness, effective pipe diameter, base demand at nodes, demand patterns of each category, etc. It could be extremely resource-consuming since the dimension of parameters could increase exponentially as the network growing larger. Provided with the network model, the demand pattern of each demand category is first checked and defined. Then seasonality adjustment factors are determined by analyzing the actual demand data. Finally, a genetic algorithm is applied to calibrate uncertain parameters such as pipe roughness, effective diameter and base demand, which are hard to be determined by simple analysis of SCADA data.

The SCADA data contains demands metered readings from 82 AMRs in Area C, such data of 2018 is used for defining patterns of residential, commercial and industrial users. Among all 82 AMRs, 71 AMRs' data are used for defining residential demand pattern, three for defining commercial pattern, and four for defining unique demand pattern of each industrial user. The multiplier at each time step is calculated as $\text{Multiplier}_{c,t} = \frac{\sum_{n=1}^{N} Demand_{c,n,t}}{\sum_{t=1}^{T}\sum_{n=1}^{N} Demand_{c,n,t}}$. Where $Demand_{c,n,t}$ is the measured demand of category *c* by AMR *n* at time step *t*, *N* is the total number of AMRs of category *c*, *T* is the total number of time steps.

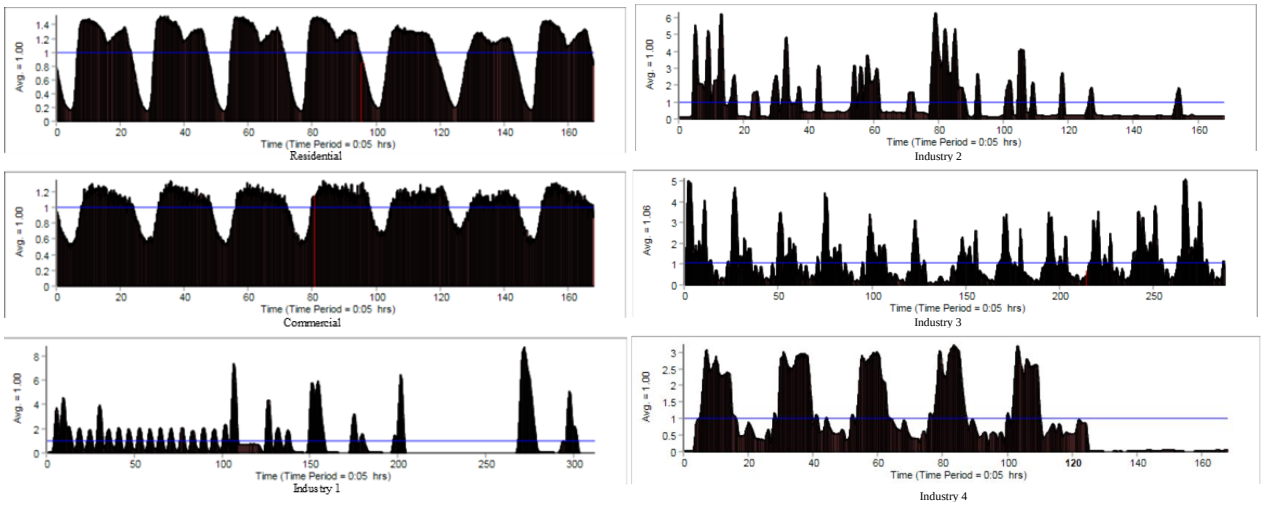


*Figure 1 Defined demand patterns for different user categories*

As shown in Figure 1, defined commercial and residential demand patterns basically follow a daily routine. Four industrial users have their own patterns, Industry 2 and Industry 4 have a seven-day pattern, Industry 1 is better described by a 13-day pattern, and Industry 3 is described by a 12-day pattern. All defined patterns have a time interval of 5 minutes.

Water consumption varies not only on a daily or weekly basis, but also on a monthly basis. Seasonality of demand is an important factor to be taken into consideration. Table 1 shows the average 5-minute demand of each category in different months in 2018. Such seasonality factors will be considered respectively for simulation of different months.

*Table 1 Demand Seasonality of Each Category*

| | | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Residential** | Avg. 5-minute Demand | 151.6 | 159.8 | 167.9 | 170.4 | 178.3 | 190.1 | 195 | 187.1 | 180.2 | 171.4 | 171.1 | 160 |
| | Annual Avg. 5-minute Demand | 173.6474611 | | | | | | | | | | | |
| | Seasonality Factor | ***0.87*** | ***0.92*** | ***0.97*** | ***0.98*** | ***1.03*** | ***1.09*** | ***1.12*** | ***1.08*** | ***1.04*** | ***0.99*** | ***0.99*** | ***0.92*** |
| **Commerical** | Avg. 5-minute Demand | 218 | 209.3 | 229.1 | 239.3 | 259.3 | 253.2 | 275.5 | 259.6 | 275.5 | 271.5 | 263.1 | 236.3 |
| | Annual Avg. 5-minute Demand | 249.3677 | | | | | | | | | | | |
| | Seasonality Factor | ***0.87*** | ***0.84*** | ***0.92*** | ***0.96*** | ***1.04*** | ***1.02*** | ***1.1*** | ***1.04*** | ***1.1*** | ***1.09*** | ***1.05*** | ***0.95*** |
| **Industry 1** | Avg. 5-minute Demand | 597.8 | 645 | 630 | 648.6 | 640.2 | 656 | 732.8 | 708.4 | 610.7 | 585.8 | 549.3 | 542.8 |
| | Annual Avg. 5-minute Demand | 629.02 | | | | | | | | | | | |
| | Seasonality Factor | ***0.95*** | ***1.03*** | ***1*** | ***1.03*** | ***1.02*** | ***1.04*** | ***1.16*** | ***1.13*** | ***0.97*** | ***0.93*** | ***0.87*** | ***0.86*** |

| | | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Industry 2** | Avg. 5-minute Demand | 1340 | 1430 | 1506 | 1471 | 1512 | 1538 | 1668 | 1659 | 1368 | 1336 | 1279 | 1214 |
| | Annual Avg. 5-minute Demand | 1443.87 | | | | | | | | | | | |
| | Seasonality Factor | ***0.93*** | ***0.991*** | ***1.04*** | ***1.02*** | ***1.05*** | ***1.07*** | ***1.155*** | ***1.15*** | ***0.95*** | ***0.93*** | ***0.89*** | ***0.84*** |
| **Industry 3** | Avg. 5-minute Demand | 1698 | 1860 | 1931 | 1811 | 1763 | 2083 | 2182 | 1942 | 1826 | 1642 | 1694 | 1625 |
| | Annual Avg. 5-minute Demand | 1835.34 | | | | | | | | | | | |
| | Seasonality Factor | ***0.93*** | ***1.01*** | ***1.05*** | ***0.99*** | ***0.96*** | ***1.13*** | ***1.19*** | ***1.06*** | ***0.99*** | ***0.89*** | ***0.92*** | ***0.89*** |
| **Industry 4** | Avg. 5-minute Demand | 33.79 | 35.56 | 37.15 | 35.79 | 38.04 | 38.71 | 41.62 | 41.37 | 33.55 | 33.88 | 31.97 | 29.52 |
| | Annul Avg 5-minute Demand | 35.93 | | | | | | | | | | | |
| | Seasonality Factor | ***0.94*** | ***0.99*** | ***1.03*** | ***1*** | ***1.06*** | ***1.08*** | ***1.16*** | ***1.15*** | ***0.93*** | ***0.94*** | ***0.89*** | ***0.82*** |

While demand patterns and seasonality can be determined by analysis of AMR demand data, some model parameters remain uncertain and hard to analyze, such as pipe roughness, effective diameter and nodal base demand. As suggested by the utility company of "L-Town", all these parameters could have no more than 10% difference of the nominal values. Genetic algorithm is used to optimize this problem.

$$\textit{Variables: } C_{demand,r}, C_{demand,c}, C_{roughness}, C_D$$

$$\textit{Constraints: } C_{demand,r}, C_{demand,c}, C_{roughness} \; in \; [0.9, 1.1], \; C_D \; in \; [0.9, 1]$$

$$\textit{Objective Function: } minimize \; Loss = min \; \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}(P_{t,i}-P_{t,i,0})^2}{N \cdot T}$$

Where $C_{demand,r}$ and $C_{demand,c}$ represent demand variation coefficients of residential and commercial users; $C_{roughness}$ and $C_D$ represents variation coefficients of pipe roughness and effective diameter. $P_{t,i,0}$ is the measured pressure at sensors, $P_{t,i}$ is the calculated pressure; N and T are the total sensor number and time steps considered in calibration. The effective diameter cannot be greater than its physical size, thus $C_D$ cannot exceed 1. Other hydraulic constraints are satisfied automatically by EPANET engine.

The first two-week data of 2018 measured pressure are selected as the calibration reference. The defined patterns of each user category and demand seasonality of January are set in the hydraulic model prior to GA optimization. The results of two GA optimizations are listed as follows. Each GA optimization has 30 and 50 population respectively, and both have 30 generations. The results of Opt. 1 are used as the calibrated parameters of hydraulic model.

*Table 2 Calibration results of GA optimization*

| | $C_{demand,r}$ | $C_{demand,c}$, | $C_{roughness}$ | $C_D$ | Loss |
|---|---|---|---|---|---|
| Opt. 1 (30 pop.) | 1.082 | 1.015 | 0.983 | 1.027 | 0.0144 |
| Opt. 2 (50 pop.) | 1.074 | 1.09 | 0.997 | 1.011 | 0.0143 |

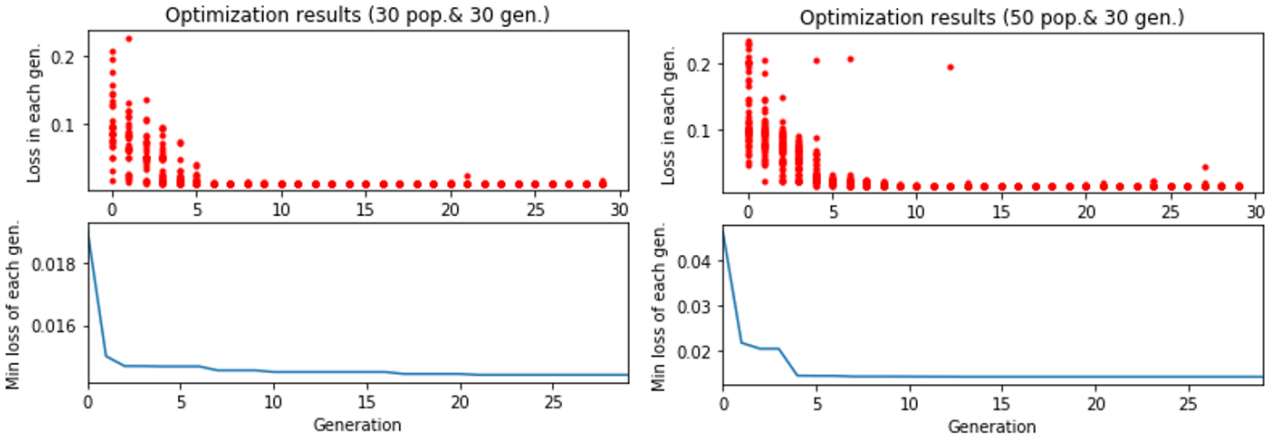


*Figure 2 Optimization losses of each generation*

## 1.2 Data Mining Classification Model

The purpose of data mining is to find the potential leaking area based on the pressure data of the network. Machine learning classification models are constructed to learn pressure change characteristics of each area, and the leakage category is identified from the pressure situation at each moment by the model. LightGBM and LSTM are selected to achieve the classification goal.

In order to build the training dataset of the classification model, labels with leakage area's sensor ID (The column number of each monitoring sensor in the table) or no leakage (marked as 0) are marked at each time step. The calibrated hydraulic model is used to simulate numerous leakage events with WNTR, ensuring every leakage condition is recorded. The simulated results and 2018 measured SCADA data are then combined as the training and testing dataset. Firstly, the difference between the pressure value $P$ and the referential value of the current monitoring point is extracted as $\Delta P_t = P_t - [0.95*(P_{max} - P_{min}) + P_{min})]$. Then, a 1-hour average pressure deficit at each moment of each sensor is extracted as the pressure feature. The sample balance of the dataset is then adjusted to ensure all types of leakage situations are learned equally by the model. At last, the dataset is scaled to a certain range with the same scale to ensure that various types of feature values have the same weight, the MinMaxScaler is used here as a scaler.

## 1.3 Leakage Isolation

Given the predicted results from data classification models, a genetic algorithm-based leakage isolation model is applied. Leakage events are simulated iteratively in EPANET and pressure results are then compared with the measured data, the most matched simulation indicates the possible leakage event in reality. The optimization problem is described as follows.

$$\textit{Variables: } LN_i \ , K_i$$

$$\textit{Constraints: } K_i \ \ in\ [0, K_{\max}],\ P_i \ > 0,\ NL_{dup} = 0$$

Where $LN_i$ is the index for leakage node $i$; $K_i$ is the emitter coefficient for leakage node $i$, which is used to simulate leakages in EPANET; $K_{max}$ is the maximum emitter coefficient; $P_i$ is the pressure at leakage node $i$; $NL_{dup}$ is the number of the duplicated nodes identified as leakages by one simulation,

$$\textit{Objective Function: } minimize\ F(LN_i \ , K_i \ ) = \ min \ \sum_{t=1}^{T} \frac{\sum_{n=1}^{N} [\frac{H^S_{n(t)} - H^O_{n(t)}}{H^O_{n(t)}}]^2}{N}$$

where $H^O_{n(t)}$ and $H^S_{n(t)}$ are the observed head and simulated head at junction $n$ at time step $t$; $N$ is the number of observed head.

A potential leakage event is represented as the index of leakage node with a positive emitter coefficient. Binary code is used for encoding the two variables that the node ID is designated as a

node index and the emitter coefficient is encoded with a value between zero and the prescribed maximum value with a specified increment. By applying such scheme, the maximum number of leakage nodes is specified and does not have to be the same as the total number of nodes. The optimized number of leakage nodes, where the calculated emitter coefficients are greater than zero, should be less than the specified maximum number of leakage nodes to be detected. Otherwise, and the model should be re-run for leakage detection.

## 2. Results and Discussions

Since the data volume of hydraulic simulations is much larger than the real data volume in 2018, 20% of the 2018 data is used as the validation set, the remaining part and hydraulic simulation data are combined as the training set, ensuring that the model can learn the measured data better. The two models are trained separately, and the model parameters and model structure are adjusted and optimized according to the test results. The training and testing loss of these two models are shown in Figure 3. The accuracies of LSTM and LightGBM are 95.85% and 97.27%, respectively. The predicted leakage time periods and leakage area's sensor ID of possible leakage events in 2018 are list in Table 3, these results are used for leakage isolation algorithm.

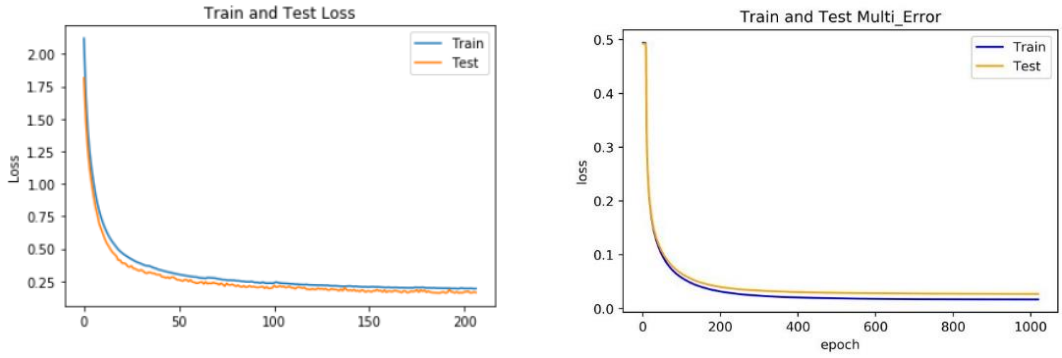


*Figure 3 Model loss results of LSTM (left) and LightGBM (right)*

*Table 3 2018 Real Leakage Repair Time VS. Predicted Leakage Time*

| Leakage area sensor ID | Leakage Repair Time (Reported) | Leakage Repair Time (Predict) | Leakage Start Time (Predict) |
|---|---|---|---|
| n144 | 2018/4/2 11:40 | 2018/4/2 11:50 | 2018/3/5.7:00 |
| n31 | 2018/8/12 17:30 | 2018/8/12 17:45 | 2018/7/18 9:00 |
| n188 | 2018/5/29 21:20 | 2018/6/1 19:00 | 2018/5/11 7:00 |
| n229 | 2018/3/23 10:25 | 2018/3/23 10:25 | 2018/3/5 7:00 |
| n296 | 2018/6/12 3:00 | 2018/6/12 3:10 | 2018/6/2 6:05 |
| n410 | 2018/11/8 20:25 | 2018/11/8 20:40 | 2018/10/23 13:35 |
| n458 | 2018/2/10 9:20 | 2018/2/10 9:25 | 2018/2/1 7:00 |
| n506 | 2018/6/2 6:05 | 2018/6/2 5:45 | 2018/5/29 21:20 |
| n644 | 2018/10/23 13:30 | 2018/10/23 14:05 | 2018/10/05 19:00 |
| n752 | 2018/9/1 17:10 | 2018/9/1 17:20 | 2018/8/12 17:30 |

Given the predicted leakage time periods and area, leakage isolation algorithm is applied. The results of 2018 are shown in Table 4 and Figure 4. It is shown that each leakage event is well-isolated as the detected leaking node locates closely to reported leaking pipe.

*Table 4 The results of leakage isolation for 2018 dataset*

| Leakage event | Predicted leakage start time | Reported leakage repair time | Reported leakage pipe | Leakage isolation algorithm results |
|---|---|---|---|---|
| 1 | 2018/2/1 7:00 | 2018/2/10 9:20 | p232 | n464 |
| 2 | 2018/3/5 7:00 | 2018/3/23 10:25 | p673 | n206 |
| 3 | 2018/3/5.7:00 | 2018/4/2 11:40 | p461 | n490 |
| 4 | 2018/5/11 7:00 | 2018/5/29 21.20 | p628 | n185 |
| 5 | 2018/5/29 21:20 | 2018/6/2 6:05 | p538 | n123 |
| 6 | 2018/6/2 6:05 | 2018/6/12 3:00 | p538 | n736 |
| 7 | 2018/7/18 9:00 | 2018/8/12 17:30 | p31 | n44 |
| 8 | 2018/8/12 17:30 | 2018/9/1 17:10 | p183 | n752 |
| 9 | 2018/10/05 19:00 | 2018/10/23 13:35 | p158 | n658 |
| 10 | 2018/10/23 13:35 | 2018/11/08 20:25 | p369 | n82 |


*Figure 4 The results of leakage isolation for 2018 dataset*

Applying the same leakage detection and isolation approach, the results of 2019 are shown in Table 5, the pipe ID is selected as the nearest pipe connected to the detected leaking node.

*Table 5 The results of leakage detection and isolation for 2019 dataset*

| Leakage event | Leakage area sensor ID | Leakage start time | Leakage repair time | Leakage isolation algorithm results |
|---|---|---|---|---|
| 1 | n506 | 2019/1/15 23:25 | 2019/2/1 10:00 | P523 |
| 2 | n726 | 2019/1/24 23:55 | 2019/2/7 10:00 | p829 |
| 3 | n458 | 2019/4/22 9:30 | 2019/5/3 13:40 | p430 |
| 4 | n613 | 2019/6/10 9:35 | 2019/7/18 8:10 | p902 |
| 5 | n296 | 2019/8/18 10:00 | 2019/10/1 7:05 | P173 |
| 6 | n188 | 2019/8/25 10:50 | 2019/9/14 11:00 | p133 |

**Keywords:** Leakage detection, Leakage isolation, Data mining, Hydraulic simulation, Genetic algorithm

## SUMMARY

Water loss due to leakages in water distribution network has been a major issue, it costs not only water resource and money but also brings potential water quality problems to the system. Researchers and engineers have been working on techniques detecting and isolating leakages in WDN and many approaches have been proposed and applied. Nowadays, the digitalization of urban infrastructures applies the SCADA in water distribution systems, which brings numerous data for analysis. This research proposed a leakage detection and isolation approach for WDN, which could effectively accomplish the goal. The hydraulic model is first calibrated by applying genetic algorithm and WNTR simulation. The demand pattern and monthly seasonality for each user category are calibrated and defined by analyzing AMR data. Uncertain WDN parameters calibration are then done by genetic optimization. Two data classification models (LSTM and LightGBM) are then used to learn how SCADA data and simulated data (especially pressure data) changes when leakages exist in network. The trained models show high accuracy on 2018 dataset and could predict suspicious leakage areas and time periods of certain leakages. Given this predicted information, a genetic-algorithm-based isolation method is applied to find out the exact leaking node or pipe. Iterative hydraulic leakage simulations are run by EPANET engine and the results of each simulation are compared with SCADA data. The most matched simulation indicates the simulated leakage event is of high confidence being an actual leakage in network. Such detection and isolation approach is applied in L-Town, using 2018 SCADA data and reported leakage list. All 10 reported leakages in 2018 are detected by data classification models and isolated by the isolation algorithm. The predicted leakage locations are close to the reported ones, indicating good performances of the proposed approach. The approach is then applied to detect and isolate leakages for L-Town in 2019.