Novel Leak Location Approach in Water Distribution Networks with Zone Clustering and Classification

Marcos Quiñones-Grueiro[1](✉), Cristina Verde[2], and Orestes Llanes-Santiago[1]

[1] Departamento de Automática y Computación,
Universidad Tecnológica de la Habana, Havana, Cuba
marcosqg88@gmail.com

[2] Instituto de Ingeniería, Universidad Nacional Autónoma de México,
Mexico City, Mexico

Abstract. A novel leak location approach for large-scale water distribution networks (WDNs) is discussed in this paper. The location task is formulated as a classification problem, and it is simplified by applying a clustering strategy. Data from each class are formed by measurements associated with leakages that occur within a specific zone of the WDN. A zone is defined as a set of nodes that share similar topological properties. Therefore, clustering is performed for network partitioning. Sensors are then placed within the network for maximizing leak detection coverage, and data of each class are generated by using the EPANET hydraulic simulator. The robustness of the proposal is demonstrated for different kinds of uncertainties and measurements' noise. A real-life network is used as case study with synthetically generated field data. The proposal achieves an improved performance for the different scenarios in comparison with the node location approach.

Keywords: Leak location · Clustering · Classification · Uncertainties

1 Introduction

Water preservation is important for the future of humanity. Water distribution networks are complex nonlinear dynamic systems continuously delivering drinking water to different types of consumers. Leakages cause a significant loss in the fluid transportation system as well as additional effects, such as low pressure, for final consumers. Therefore, leakage monitoring has stimulated many researchers in recent years [9,10,13,14].

Leak location approaches have been proposed by using analytic, data-driven and mixed models. Leakages are located by analyzing the difference between measurements and synthetic data generated through a model. Analytic models of the network consider the main physical laws that describe the system's operation. Data-driven models capture the network behavior from a representative

J. A. Carrasco-Ochoa et al. (Eds.): MCPR 2019, LNCS 11524, pp. 37–46, 2019.
https://doi.org/10.1007/978-3-030-21077-9_4

sample of measurements obtained over varying conditions. Mixed approaches are formalized by using an analytic model and a data-driven decision tool. The latter has shown promising results recently [9].

Since it is usually considered that leaks occur in nodes, the goal of traditional location approaches is to find the leakage node [9]. Depending on the number of sensors, there can be many indistinguishable leak signatures for different nodes because of uncertainties of the model and measurement noise. As a consequence, the location problem may require many sensors for achieving a satisfactory performance in large-scale networks. Researchers have proposed as an alternative solution forming the classes by grouping nodes into zones of the network [10,13,14]. The idea of locating leaks in predefined network zones is appealing from a practical point of view because it allows the operators to narrow down the leakage location to a bounded area. Therefore, leakage search with specialized equipment is faster because there is certainty about the zone where the leak is located.

Wachla et al. proposed a set of classifiers for locating leaks in sub areas [10]. In their work, flow meters are installed at different nodes, and areas are defined by domain experts. Zhang et al. solved the zone location problem with support vector machines by considering pressure measurements [14]. Network zones are defined according to the network variables' response to leakages in different nodes. Therefore, their results depend on the network operating conditions. A similar approach was put forward by Xie et al., but linear classifiers are combined with a sparse representation for improved results [13]. The latter research explores the influence of uncertainty in the measurements and concludes that selecting the sensors' number depends on the measurement precision for a desired location accuracy. In realistic conditions, there are several sources of uncertainty associated with the hydraulic model parameters, consumers' demand and measurements' noise. None of the previously mentioned works, however, explores the impact of all these uncertainties in the location performance. Moreover, the zone partitioning results depend on the network operating conditions and leakage sizes simulated.

The main contribution of this paper is a novel leak location approach for WDNs that combines a topological clustering strategy with classification tools. Operational zones are defined by using the k-medoids method, which considers the topological parameters of the network nodes. Pressure sensors are placed for guaranteeing the detection of small leakages. Classification tools selected for the leak location task are Random Forests (RFs) and Support Vector Machines (SVMs). Both classifiers were selected because of their successful results in many papers and their different working principles. The proposed approach allows selecting a reliable and robust leak location by setting the trade-off among the number of zones, the number of sensors and the location accuracy under realistic conditions. A real large-scale network, the Modena WDN, is considered for demonstrating the advantages of the proposal against different uncertainties in comparison with recent node location approaches.

The structure of the paper is the following. The modeling framework of the WDN is introduced in Sect. 2. The zone clustering strategy is described in Sect. 3. The classification approach is outlined in Sect. 4. For demonstrating the advantages of the proposal, the Modena WDN is introduced, and the uncertain scenarios considered are detailed in Sect. 5. The results and discussion are presented in Sect. 6. Finally, conclusions and directions for future work are proposed.

2 Water Distribution Networks

Water networks are formed by n_1 junctions and n_2 nodes spatially distributed across a geographical area. Two main physical laws that govern the behavior of demand-driven WDNs are as follows: (1) the net inflow must be equal to the net outflow for any node of the network, and (2) the sum of pressure heads around any loop of the network is equal to zero. In general, leakages are considered as extra demands that occur at existing nodes according to the following equation

$$\sum_{ni=1}^{Ni} q_{ni}(t) = d_i(t) + l_i(t); \quad l_i(t) = C_e h_i(t)^{\gamma} \tag{1}$$

where $h_i(t)$ is the pressure head, l_i is the leakage outflow, d_i is the total demand, Ni is the number of branches connected to the node i, $q_{ni}(t)$ denotes the flow of the branch ni, C_e is the emitter coefficient size and $\gamma = 0.5$ [7]. To distinguish a leak from a demand deviation, some properties for the demand must be known. Therefore, leakage location is generally performed by monitoring flows or pressure heads during minimum night flow conditions because the demand behavior is easy to characterize.

3 Zone Clustering

The zone partitioning for WDNs is performed for many purposes. In particular, it is commonly formulated to establish district metered areas (DMAs). Clustering is usually applied to define the shape and dimension of network zones. Given a data set of topological parameters $D = \{\mathbf{c}\}_{i=1}^{n}$ of n nodes with $\mathbf{c} \in \Re^m$, the clustering task can be formulated as finding the z clusters of nodes $G_{j=1}^{z} = \mathbf{c}_1, ..., \mathbf{c}_{n_j}$ that maximize/minimize an optimization function. The three main variables considered for the $\mathbf{c}$ vector are the geographical coordinates (X,Y) and the topological height of each node. Different methods have been applied for zone division in WDNs [5]. Since uniformity is the main concern within the scope of this paper, the k-medoids clustering algorithm will be used. The optimization problem is the following

$$\mathbf{min}(\mathit{SSE(G)}) = \sum_{i=1}^{z} \sum_{j=1}^{n_j} \|\mathbf{c}_{n_j} - \mathbf{c}^*{}_i\|^2 \tag{2}$$

where $\mathbf{c^*}_i$ is the vector of parameters of a specific node and $\sum_{j=1}^{z} |G_j| = n$. Partitioning around medoids is the algorithm used in this paper for solving this optimization problem. Further details can be found in [4].

Network partitioning can be performed by using hydraulic and topological indicators to form the $\mathbf{c}$ vector. Hydraulic indicators require a hydraulic model of the network [5]. Topological indicators are normally easy for computation and use. Moreover, the shape of the network partitions obtained by employing topological indicators does not depend on the simulated network operating conditions. Therefore, the topological parameters considered for each node are its coordinates and elevation. Hence, nodes are grouped together according to their geographical locations. There are many indicators used for assessing the quality of network partitioning algorithms. Nonetheless, the criterion considered in this paper is the uniformity: a similar number of nodes throughout all the clusters. In addition, it is recommended to visually evaluate the results since the partitioning and the linkage among clusters can be analyzed intuitively [6].

4 Classification Tools

A mixed model/data-driven leak location strategy is described next. The estimated consumer demands $\tilde{\mathbf{x}} = \tilde{\mathbf{d}} \in \Re^N$ are used for generating the system's response $\tilde{\mathbf{y}} \in \Re^p$ by using the nominal analytic hydraulic model. The measured variables from the real network can be flows $\mathbf{q} \in \Re^{n_1}$ in n_1 pipes and pressure $\mathbf{h} \in \Re^{n_2}$ at n_2 nodes such that $\mathbf{y} = [\mathbf{q}, \mathbf{h}] \in \Re^{p=n_1+n_2}$. It is considered that single leaks $\Omega = \{\mathbf{l_1}, \mathbf{l_2}, ..., \mathbf{l_z}\}$ can occur at any of the z network nodes. Thus, a residual vector $\mathbf{r} \in \Re^p$ (with p as the number of sensors installed in the network) provides a leakage signature according to the leakage location. From a pattern recognition point of view, the classification task consists of mapping the feature space ($\mathbf{r}$) onto a set of z classes (leak location $\tilde{\mathbf{l}}_\mathbf{i}$) by using a decision function: $g(\mathbf{r})\colon \Re^p \rightarrow \Omega$. The parameters of $g(\mathbf{r})$ are then estimated off-line by sampling from the classes population according to the *learning from examples* paradigm [3]. Two classification tools selected in this paper are described next.

4.1 Random Forests

Random forests is a machine learning method used for classification and regression [1]. In the former task, an ensemble of decision trees is used for making the class decision. A single decision tree is a recursive and partition-based classifier. This classifier splits the data space into regions by using axis-parallel hyperplanes

$$g(\mathbf{r}) = \mathbf{w}^T \mathbf{r} + b \tag{3}$$

where $\mathbf{w}$ and b (bias) are used to define the hyperplane position; and $\mathbf{r} \in \mathbb{R}^{\mathrm{p}}$ denotes a measurement vector. The value of $\mathbf{w}$ is restricted *a priori* to one of the standard basis vectors $\mathbf{e_1}, ..., \mathbf{e_p}$, where $\mathbf{e_1} \in \mathbb{R}^{\mathrm{p}}$ has a 1 for the j dimension and 0 for the others. A hyperplane specifies a decision or split point. The selection of split points is made in this work by minimizing the measure known as the Gini diversity index.

The ensemble is formed by a number of decision trees built by applying two preprocessing operations on the original data set: bootstrapping and random feature selection. The former consists of generating training sets by randomly sampling with replacement from the original data set. The latter is randomly selecting a limited number of features at each node when building the tree without pruning. Once a large number of trees are built, new data are classified by aggregating the outputs of all trees by applying a majority voting strategy. While individual decision trees tend to overfit, random forests present a good generalization performance thanks to the previous two operations. The number of variables randomly selected for each tree in this work is the square root of the number of variables.

4.2 Support Vector Machines

The objective behind the support vector machine method is defining the optimal separating hyperplane that maximizes the margin w among the closest observations of two different classes that form a data set. These observations are called support vectors [8]. The separating hyperplane $g(\mathbf{r})$ of two classes is defined with Eq. (3), but the values of w are not restricted as in decision trees. Conversely, w and b are defined by solving the following dual optimization problem

$$\max \; W(\mathbf{a}) = \left(\sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i,j=0}^{m} a_i a_j g_i g_j \; K(\mathbf{r_i}, \mathbf{r_j}) \right)$$

$$\text{subject to } \sum_{i} g_i a_i = 0; \quad 0 \leq a_i \leq C \tag{4}$$

where C represents the error penalty, $\mathbf{a} \in \Re^m$ are the Lagrange multipliers, m is the number of training examples that form the data set $X \in \{\mathbf{r_i}, \mathbf{y_i}\}^m$ with the label vector $\mathbf{y} \in \{1, -1\}$, and $\mathrm{K}(\mathbf{r_i}, \mathbf{r_j})$ is a kernel function that allows access to spaces of higher dimensions. The Radial Basis Function kernel is selected in this work because of its generality and successful results. The extension of this method to multi-class classification problems is developed by applying discriminant strategies. The one-against-one approach is selected in this work.

5 Case Study: Modena Network

The Modena network is a reduced version of the WDN of the medium-sized Italian city. It is formed by 317 pipes and 268 demand nodes with a required minimum pressure head of 20 m. The network is gravity-fed by four reservoirs, and it is completely looped as shown in Fig. 1. The pipe diameters are set according to [11]. Pressure head sensors are used here because they are cheaper and easier to install and maintain than flow meters. The location of the sensors is selected by maximizing the leak detection coverage. The *Darwin Sampler* tool was used for placing a specific number of sensors [12]. Single leakage events are generated by considering the minimum leak size that is desired to be located. This occurs for an emitter coefficient at each node with a magnitude of 0.1. The sensitivity selected for pressure head sensors is 0.01 m. The number of leakage scenarios is selected as three times the number of network nodes (1000 leakage scenarios) according to the software's recommendation.

5.1 Uncertainty Simulation for Realistic Scenarios

A steady-state simulation of the network is performed with the package EPANET [7] coupled with $MATLAB^{©}$. A sampling period of 15 min is considered, and hourly average values of the measurements are used for leak location, which aims to reduce the uncertainty effect [9]. A total of 120 samples (hourly averages) are generated for each node by considering minimum night flow conditions, but the final data set is formed by grouping the nodes' data corresponding to each class according to the zones' distribution. The uncertainty effects are obtained by using the following equation

$$\theta_r = \theta_t + \theta_u \tag{5}$$

where θ_r represents the uncertain parameter and θ_t and θ_u are the true and the added uncertainty, respectively. All values of θ_u are generated from a uniformly sampled distribution. The following unknown disturbances are all simulated for resembling real conditions

1. Leak size variability. Uncertainty is related to the emitter coefficient size that is considered within the range $C_e \in [0.1, 2]$. The outflow of leakages is between 0.5 lps and 12 lps (approximately 0.1% to 3% of the network's total demand).
2. Measurements uncertainty. Measurements are corrupted with 5% noise amplitude.
3. Pipe roughness uncertainty. Hazen-Williams coefficient (CHW) uncertainty is simulated for $CHW \in [125, 130]$.
4. Estimated demand uncertainty. An uncertain demand is considered with 10% amplitude around the nominal consumption of each node.

6 Results and Discussion

Performance measures for classification problems are usually calculated by using a confusion matrix $A = [A(i,j)]$. The element (i,j) of A represents the observations with the true class label i, which are classified as class j. Location performance is estimated by considering the identification of leaky nodes within the specific predefined zone where they belong. Thus, the percentage of data that has been correctly classified determines the overall accuracy (Ac). Given z zones, Ac is computed according to $Ac = \frac{1}{m}\sum_{i=1}^{z} A(i,i)$ where m is the number of observations.

The parameters of the classifiers are set by using 10-fold cross validation. The accuracy displayed in the figures is estimated by using the test data that have not been used for adjusting the parameters. For the RFs classifier, the number of trees is adjusted to 100 by analyzing the out-of-bag error improvement. This large number of trees may take a long time to prepare off-line, but the classifier will not overfit [1]. For the SVM classifier, the parameters $\{C, \sigma\}$ were adjusted for each scenario by using a grid search for the interval $C \in 2^{\eta}$, $\eta \in [-2, 5]$ and $\gamma \in 2^{\eta}$, $\eta \in [-5, 3]$. The LIBSVM library was used for this purpose [2]. Since minimum night flow (MNF) conditions usually last for six hours (12 a.m. to 6 a.m.), the Bayes rule can be applied to the probable leak locations throughout a time window of up to six observations to obtain the leak location decision [9].

Zone clustering results for 5 and 25 zones are presented in Fig. 1. As it is observed, the network partitioning is reasonable from a practical point of view. The number of nodes per zone for the clustering of 5 and 25 zones is observed in Fig. 2. When the number of zones increases, the uniformity of the node distribution improves. The performance of the proposed approach depends on two elements: the number of pressure sensors and the number of zones. The desired result is a satisfactory accuracy of over 90% to guarantee the reliability of the location method for the network operators. It is useless to implement a method with poor performance because operators will ignore its results in the long term. There is a trade-off among the sensors, zones and accuracy that is shown in Fig. 3.

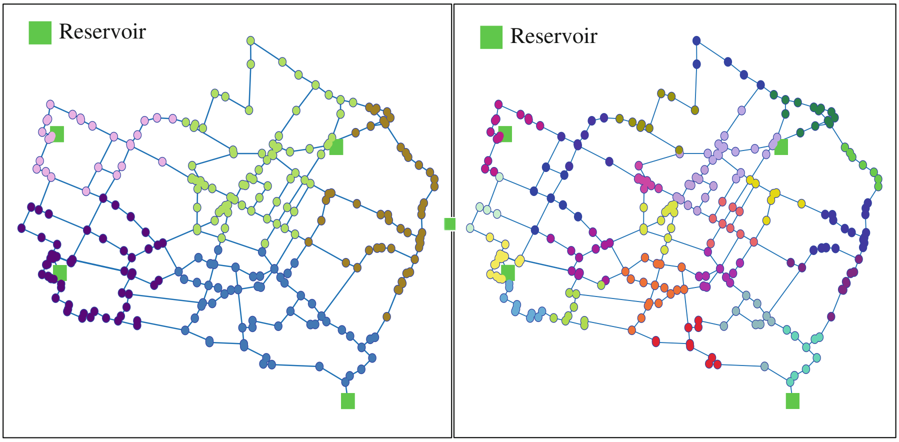

Fig. 1. Zone clustering results with topological parameters and the k-medoids method for 5 (left) and 25 zones (right)

Leak location performance degrades depending on the number of pressure sensors available. When 5 sensors are placed, only 5 leakage zones can be distinguished with a satisfactory performance by using both classifiers. When the number of sensors increases to 10, up to 15 leakage zones can then be isolated

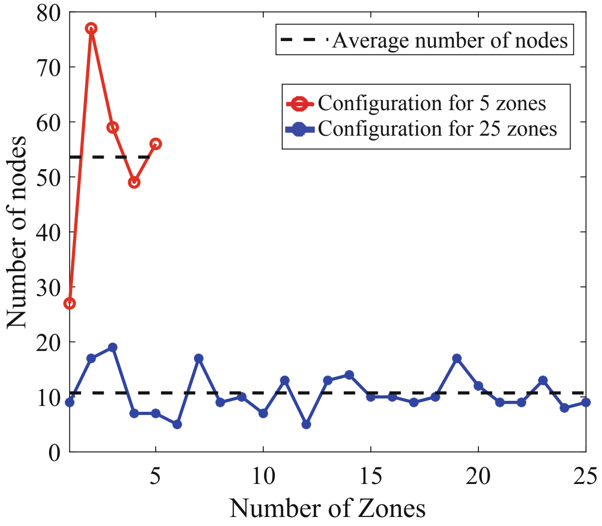

Fig. 2. Node uniformity with respect to the number of zones

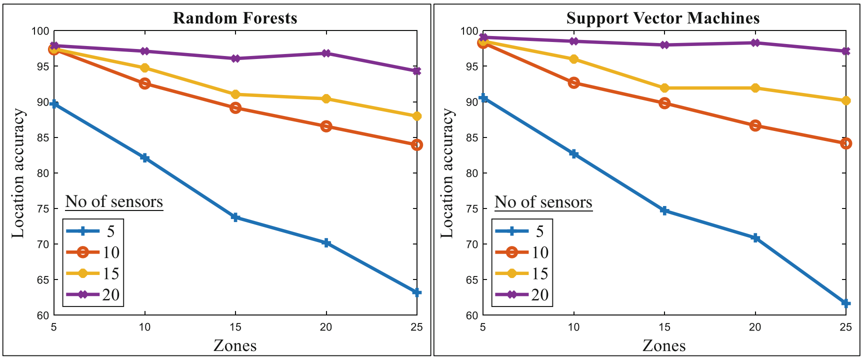

Fig. 3. Leak location performance for different configurations of sensors and zones with a time horizon of three observations (three hours)

with an overall testing accuracy of 90%. Such results confirm the assumption that clustering the nodes into multiple zones simplifies the classification problem when the number of classes is elevated.

To compare the results obtained by using the proposal presented in this paper with the node location approach proposed in [9], some simulations are developed by employing the SVM classifier. The leak location results are shown in Fig. 4. For 5 sensors only 58% accuracy can be achieved by considering a time horizon of 24 observations. Even for 25 sensors, the top performance is 83% for 24 observations. This implies requiring data from four days (considering that minimum night flow conditions last six hours in the best case) for making a decision. With the proposal only 10 sensors are required for making a decision with 90% location accuracy. Therefore, it presents a superior reliability and lower cost. Moreover, since only data from one day are required, the location decision is performed in less time than with the node location approach where data from four days are necessary.

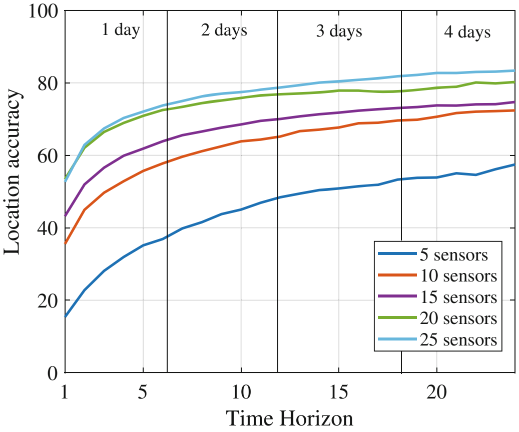

Fig. 4. Leak node location performance for a varying time horizon under MNF conditions by using the SVM classifier with the approach proposed in [9]

7 Conclusions

In this paper, a novel leak location method based on network zones is presented. The proposal uses a topological clustering strategy to divide the WDN in logical zones from a practical point of view and combines this clustering strategy with classification tools for obtaining a satisfactory performance in the leak location against uncertainties. The proposal allows establishing an adequate relation between the number of zones and the number of sensors to be used in the leak location assessment. The simulated experiments with a real network demonstrate that it is possible to obtain a reliable leak location with fewer sensors and within a shorter time horizon than other recently proposed leak location methods. The latter represents superior reliability and lower costs in the leak location task.

Acknowledgments. This work is supported by IT100519-DGAPA-UNAM and 280170conv2016-3 SENER-CONACYT, Fondo Sectorial CONACyT-Secretaría de Energía-Hidrocarburos.

References

1. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
2. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 1–27 (2011)
3. Heijden, F.V.D., Duin, R., Ridder, D.D., Tax, D.: Classification, Parameter Estimation and State Estimation. Wiley, Hoboken (2004)
4. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, Hoboken (2009)
5. Liu, H., Zhao, M., Zhang, C., Fu, G.: Comparing topological partitioning methods for district metered areas in the water distribution network. Water **10**(4), 368 (2018)
6. Perelman, L.S., Allen, M., Preis, A., Iqbal, M., Whittle, A.J.: Automated subzoning of water distribution systems. Environ. Model. Softw. **65**, 1–14 (2015)
7. Rossman, L.A.: Water supply and water resources division. National Risk Management Research Laboratory. Epanet 2 User's Manual. Technical report, United States Environmental Protection Agency (2000). http://www.epa.gov/nrmrl/wswrd/dw/%0Aepanet.html
8. Scholkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization Optimization, and Beyond. MIT Press, Cambridge (2002)
9. Soldevila, A., Fernandez-canti, R.M., Blesa, J., Tornil-sin, S., Puig, V.: Leak localization in water distribution networks using Bayesian classifiers. J. Process Control **55**, 1–9 (2017)
10. Wachla, D., Przystalka, P., Moczulski, W., Wachla, D., Przystalka, P., Moczulski, W.: A method of leakage location in water distribution networks using artificial neuro-fuzzy system. IFAC-PapersOnLine **48**(21), 1216–1223 (2015)
11. Wang, Q., Guidolin, M., Savic, D., Kapelan, Z.: Two-objective design of benchmark problems of a water distribution system via MOEAs: towards the best-known approximation of the true pareto front. J. Water Resour. Plan. Manag. **141**(3), 1–14 (2015)
12. Wu, Z., Wang, Q., Butala, S., Mi, T., Song, Y.: Darwin optimization user manual. Technical report, Bentley Systems, Incorporated, Applied Research Group (2012)
13. Xie, X., Hou, D., Tang, X., Zhang, H.: Leakage identification in water distribution networks with error tolerance capability. Water Resour. Manag. **33**(3), 1233–1247 (2019)
14. Zhang, Q., Wu, Z.Y., Zhao, M., Qi, J.: Leakage zone identification in large-scale water distribution systems using multiclass support vector machines. J. Water Resour. Plan. Manag. **142**(11), 04016042 (2016)