Cost-effective sensors placement and leak localization – the Neptun pilot of the ICeWater project

Antonio Candelieri, Davide Soldi and Francesco Archetti

ABSTRACT

This paper extends previous research to analytically identify leaks within a water distribution network (WDN), by combining hydraulic simulation and network science based data analysis techniques. The WDN model is used to run several ‘leakage scenarios’, by varying leak location (pipe) and severity, and to build a dataset with corresponding variations in pressure and flow, induced by the leak. All junctions and pipes are considered for potential pressure and flow sensors deployment; a clustering procedure on these locations identifies the most relevant nodes and pipes, and cost-effectiveness was considered. A graph is then generated from the dataset, having scenarios as nodes and edges weighted by the similarity between each pair of nodes (scenarios), in terms of pressure and flow variation due to the leak. Spectral clustering groups together similar scenarios in the eigen-space spanned by the most relevant eigen-vectors of the Laplacian matrix of the graph. This method uses superior traditional techniques. Finally, support vector machines classification learning is used to learn the relation between variations in pressure and flow at the deployed meters and the most probable set of pipes affected by the leak.

Key words | leakage localization, leakage management, spectral clustering, support vector machines, water distribution networks

INTRODUCTION

At the present time, innovative and smarter management of urban water distribution networks (WDNs) is needed to achieve higher levels of efficiency. This end can be achieved by exploiting the huge amount of data already generated by and stored in the information communication technology (ICT) systems adopted in WDN management operations, such as supervisory and control data acquisition systems and hydraulic simulation, as well as new data made available by the introduction of smart metering solutions (automatic meter reading). The availability of data, exploitation of simulation software and recent research on data analysis techniques for supporting decision-making will enable a shift toward a smart water management paradigm.

As reported by Alegre et al. (2006), the International Water Association highlighted the relevance of improving the leakage management process and also provided some specific performance indicators. Puust et al. (2010) proposed a formalization of the leakage management process, consisting of three different phases: assessment, detection and physical localization. Worldwide, urban WDNs suffer leakage, mainly due to the age of the existing infrastructures, implying service failures or disruptions, large amounts of non-revenue water, increasing costs for energy and rehabilitation, in spite of ever-tightening budgetary constraints.

Several approaches for analytically localizing leaks in a WDN have been proposed, mostly based on the idea that leaks can be detected by correlating actual modifications in flow and pressure within the WDN to the output of a simulation model whose parameters are set to evaluate the effect induced by a leak in a specific location and with a specific severity. Approaches based on machine learning, statistics, and probabilistic modeling have been investigated (Caputo & Pelagagge 2003; Poulakis et al. 2003; Xia et al. 2006; Sivapragasam et al. 2007; Behzadian et al. 2009; Xia & Guo-Jin 2010; Nasir et al. 2010; Lijuan et al. 2012).
While most of these approaches focused on localizing leaks in pipes, another recent and relevant proposal involves a combination of hydraulic simulation and support vector machine (SVM) classification to identify leaks at junctions according to the pressure and flow values (Mashford et al. 2012).

With respect to machine learning, another relevant research field supporting more effective and efficient leakage management is focused on detection of leaks and bursts through the analysis of real-time sensors data, without using any simulation (Romano et al. 2011, 2013). However, detection and localization occur in different phases of the leakage management process: the former is aimed at identifying if and when a leak exists, the latter is aimed at inferring a possibly restricted set of pipes, probably leaky, to be physically checked, reducing time and costs for investigations and intervention.

This paper builds on previous research in the definition and application of a reliable approach for localizing leaks analytically (Candelieri & Messina 2012; Candelieri et al. 2013a, b).

We present: (i) a deeper investigation of the benefits provided by network-science based machine learning approaches (i.e. spectral clustering) with respect to more ‘traditional’ ones; (ii) the proposition of a method to support a cost-effective placement of flow and pressure meters, with respect to reliability of analytical leakage localization; and (iii) the identification, through SVMs, of a reliable relationship between the modifications in pressure and flow (at the monitoring points) and the set of pipes most probably affected by a leak.

All the results are in a related hydraulic simulation model related to a real district meter area (DMA) of the WDN in Timisoara, Romania, one of the two pilots of the FP7-ICT project IceWater (ICT solutions for efficient water resources management, www.icewater-project.eu) co-funded by the European Commission.

**MATERIALS AND METHODS**

**Description of the pilot**

Figure 1 depicts the IceWater pilot (Neptun) in Timisoara, Romania. The pipe infrastructure of this DMA is about 4,000 m long, overall. Pipe lengths range from 0.034 to 83.808 m (12.778 ± 13.156 m), and pipe diameters range from 50 to 500 mm (108.0 ± 76.5 mm) and roughness (unitless) varies from 1 to 150 (111 ± 16). No tanks are installed, only one reservoir is present and all the control elements (27 valves) are opened.

With respect to the demand, which is used as input to run the simulation, 92 of the overall 335 junctions (nodes) are used to represent customers. Overall, the base demand varies from 0.00038 to 0.1601 L/s (average 0.0385 L/s, standard deviation 0.0411 L/s). Since no smart meters were available at the time of the study, the daily pattern used to model the base demand is based on historical water demand data at the inlet point (aggregated demand) of the DMA. Therefore, while the distribution in space is quite accurate – based on billing data – the distribution in time is more approximate; this limitation could be overcome by installing smart metering devices.

**Hydraulic simulation of leakage scenarios through EPANET**

The hydraulic simulation software EPANET (downloadable free from the US Environmental Protection Agency website: http://www.epa.gov/nrmrl/wswrd/dw/epanet.html) is a widely used tool for modeling WDNs, performing what-if scenario simulation and running optimization algorithms for supporting decisions both at operational, planning and strategic level.
In the proposed approach, EPANET is used to simulate a wide set of leakage scenarios, consisting of placing, in turn, a leak in each pipe and varying its severity in a given range (0.015–0.1,575 L/s). The range used has been identified according to the variation in the discharge coefficient – in the pressure-dependent model of leakage – usually adopted in the current literature in order to develop approaches for leakage detection and localization. At the end of each run, EPANET provides pressure and flow values at each junction and pipe, respectively. Then, variations in pressure and flow, induced by the leak, are computed with respect to the simulation of the faultless network model. Results obtained are stored in a dataset, together with the information related to the affected pipe and the damage severity.

More details about the leakage scenarios generation process as well as the pressure-dependent leak modeling have been described previously (Candelieri & Messina 2012).

**Clustering leakage scenarios and quality measure**

Clustering leakage scenarios previously generated through EPANET are the core of the proposed approach. The aim is to group together scenarios (rows of the dataset) that are similar in terms of variations in pressure and flow induced by the corresponding leak. Only information on pressure and flow (that is the effect of the leak) is taken into account during this process, while information on leaky pipes and leak severity is ignored. Several clustering algorithms are available; all of them need a specific measure (distance or similarity) to be defined in order to compare two objects, that in this case are two vectors of pressure and flow variations at junctions and pipes.

At the end of the clustering process, a measure should be adopted to evaluate the quality of the identified clusters. This measure has to be different from the one used to perform clustering and enable an evaluation of the quality of the solution with respect to goal. Although several measures have been proposed to evaluate the validity of clustering procedures, evaluating the capability to localize leaks has required the definition of a specific measure, namely the previously proposed ‘localization index’ (Candelieri et al. 2013a). The localization index (LI) for each cluster (LI\(k\)) requires retrieval of the information on leaky pipes related to each scenario and is then computed as the number of distinct pipes related to the scenarios in that cluster with respect to the overall number of pipes in the WDN:

\[
LI_k = \frac{|\text{pipes} - |\text{pipes}_k|}{|\text{pipes}| - 1}
\]

(1)

where |pipes| is the overall number of pipes of the WDN and |pipes\(k\)| is the number of leaky pipes of the scenarios into cluster \(k\).

The maximum value (LI\(_k\) = 1) is obtained when the cluster \(k\) contains scenarios all related to leaks simulated only in one pipe (i.e. |pipes\(k\)| = 1). The minimum value (LI\(_k\) = 0) is obtained if the cluster \(k\) contains scenarios related to all the pipes of the WDN (i.e. |pipes\(k\)| = |pipes|).

Previously, we computed the overall LI as the simple average of LI\(_k\), but in this case the average has been weighted by the number of distinct pipes in each cluster:

\[
LI = \frac{\sum_{k=1}^{K} LI_k |\text{pipes}_k|}{\sum_{k=1}^{K} |\text{pipes}_k|}
\]

(2)

where \(K\) is the overall number of clusters.

We now propose another relevant measure to evaluate how much a clustering algorithm is able to put in the same cluster scenarios related to the same (leaky) pipe and to all the different severity values. This index is the ‘quality of localization’ (QL) and is defined, for each cluster \(k\), as:

\[
QL_k = \frac{\sum_{j\in S} \frac{n_p^j}{|S|}}{n_p^k}
\]

(3)

where \(S\) is the set of different severities used (and |\(S\)| is the overall number of severity values), \(p_j\) is a pipe having a leak whose severity is the \(j\)th in \(S\), \(n_p^j\) is the number of scenarios in cluster \(k\) associated with leaks with the \(j\)th severity and \(n_p^k\) is the number of distinct pipes related to the scenarios in cluster \(k\).

The maximum value of QL\(_k\) = 1, obtained when in the cluster \(k\) are all the scenarios related to all the severity values of the correspondent leaky pipes.

The overall QL for a clustering algorithms is given by the average of QL\(_k\).
Finally, a global index $LI^*$ is defined, combining $LI$ and $QL$:

$$LI^* = LI \times QL$$

### Spectral clustering

Although spectral clustering (Jaakkola 2006; Luxburg 2007) has been proposed in order to solve graph clustering problems (Schaeffer 2007), it usually outperforms traditional clustering algorithms, such as the $K$-means or other partitioning algorithms, when applied to not-relational data points datasets.

The final goal is the same both for traditional and spectral clustering, i.e. partitioning objects (leakage scenarios in this case) into subsets so that objects in a cluster would be more similar than outside the cluster.

However, graph clustering strategies, such as spectral clustering, solve the problem by taking into account a graph-based structure of the relations (edges) between objects (nodes). The aim is to group nodes of the graph into sub-graphs (clusters), maximizing the sum of the weights on the edges within each cluster (intra-cluster similarity) while minimizing the sum of the weights on the edges connecting nodes in different clusters (inter-cluster similarity).

In this study, nodes are leakage scenarios and edges are weighted by the similarity between two scenarios, computed as triangle-similarity (Zhang et al. 2011) between the two correspondent vectors of variations in pressure and flow at the monitoring points; the resulting network structure is a similarity graph between leakage scenarios.

The solution of the graph clustering problem can be easily described in the case of bi-partitioning. Given two sets of nodes (clusters), $C_1$ and $C_2$, the objective is to minimize:

$$\text{cut}(C_1, C_2) = \sum_{x_i \in C_1, x_j \in C_2} s_{ij}$$

where $s_{ij}$ is the weight of the edge $(x_i, x_j)$; usually this weight represents the similarity ($s$) between the two nodes.

A $n$-dimensional vector $p$ ($n$ is the number of nodes in the graph) is used to represent the association of each node to cluster $C_1$ or $C_2$:

$$p_i = \begin{cases} +1 & \text{if } x_i \in C_1 \\ -1 & \text{if } x_i \in C_2 \end{cases}$$

The graph clustering problem can be formulated as minimization of the following function $f(p)$:

$$f(p) = \sum_{x_i, x_j \in V} L_{ij}(p_i - p_j)^2 = p^T Lp$$

where $L_{ij}$ are the entries of the Laplacian matrix, the core of spectral clustering. Alternative definitions have been proposed and studied through graph theory (Chung 1997); the usual definition is:

$$L = D - A$$

where $A$ is the affinity matrix of the undirected graph and $D$ is the degree matrix, with each entry defined as:

$$d_{ii} = \sum_j a_{ij}, i = j$$

$$d_{ij} = 0, i \neq j$$

The most important properties of the $L$ matrix are:

- it is symmetric and positive semi-definite (it has $n$ non-negative, real-valued eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, irrespectively of their multiplicity);
- its smallest eigenvalue is 0 (where its multiplicity indicates the number of distinct connected components).

Many applications use a normalized Laplacian matrix instead of the basic one; the most common definition for the normalized Laplacian matrix is the following:

$$L_{\text{norm}} = I - D^{-1/2}AD^{-1/2}$$

The combinatorial complexity of minimizing (7) can be prohibitive for real-world networks (Ng et al. 2001; Luxburg 2007). However, a simple algebraic solution to the problem
was proposed in (Fiedler 1973): in particular, he used the result of the Rayleigh theorem and identified the second smallest eigen-vector of the Laplacian matrix (usually known as Fiedler vector) as the vector $p$ which provides the optimal bi-partitioning of the graph.

This result has permitted implementation of recursive bi-partitioning spectral clustering approaches (Hagen & Kahng 1992) in order to perform partitioning in $K > 2$ groups. However, this approach requires the computation of matrices and eigenvalues, as well as the use of the Fielder vector, for each sub-graph until the desired number of clusters is reached.

Another possible schema to solve the $K$-partitioning uses a data representation in the – usually low-dimensional – space of relevant eigen-vectors (Ng et al. 2001; Luxburg 2007). The relevant eigen-vectors are the first $l$ smallest: the $l$-th eigenvalue is the one showing a sufficiently large variation in the eigengap, i.e. the difference between two successive eigenvalues in the list of eigenvalues sorted in ascending order.

For example, the $K$-partitioning approach proposed in Shi & Malik (2000), consists of selecting the $l$ smallest non-zero eigenvalues and performing a traditional $k$-means clustering on the resulting dataset having $n$ rows (nodes of the graph) and $l$ columns (eigen-vectors corresponding to the $l$ smallest eigenvalues). Any other traditional clustering algorithm may be applied in the eigenspace.

**Identifying leaky pipes through SVMs**

After clustering the leakage scenarios, the next step consists of discovering a reliable relation between the variations in pressure and flow, due to a leak, and the corresponding cluster, which permits retrieval of the set of corresponding leaky pipes. This relation allows the time for investigations and rehabilitation to be decreased: when a leak is detected, e.g. with traditional methods, such as minimum night flow analysis (Liemberger & Farley 2004; Behzadian et al. 2009; Izquierdo et al. 2011), the actual pressure and flow measurements at the monitoring points are compared with those obtained through simulation of the faultless network model, in order to compute the variations in pressure and flow and finally identify only a restricted number of pipes to check physically.

In our previous study (Candelieri & Messina 2012), we proposed comparison of the obtained vector of values with those related to the centroids of the clusters, and selected the cluster associated with the most similar centroid. However, since the spectral clustering procedure implicitly applies a non-linear transformation from the input space (related to the variations in pressure and flow) to the eigen-space spanned by the most relevant eigen-vectors of the Laplacian matrix, similarity computed in the input space does not guarantee that the association of the computed vectors to a specific cluster is correct.

To improve the reliability of the localization process, a SVM classifier has been trained taking the variations in pressure and flow of each scenario as input and the correspondent cluster provided by spectral clustering as target output (class label). Thus, the SVM classifier learns to approximate the non-linear mapping performed by spectral clustering and to estimate the most probable cluster which an actual vector of variations in pressure and flow belongs to.

Figure 2 shows the overall workflow, presenting the mapping performed by the spectral clustering. The SVM permits application of spectral clustering only to a more restricted, even if relevant, set of leakage scenarios, requiring a smaller scenarios graph, reducing complexity in memory and time, and guaranteeing adoption of a reliable approximation of spectral clustering to identify the corresponding scenarios cluster for any new vector of variations in flow and pressure.

**Support vector machine classification**

The basic idea of SVMs (Vapnik 1998) consists of searching for a hyper-plane to optimally separate instances (represented as vectors) belonging to different classes and, at the same time, maximizing the distance of the instances from the hyper-plane.

This formulation is usually known as hard margin SVM. Given a dataset $D$ of $n$ instances, it can be represented as:

$$D = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\} \text{ with } i = 1, \ldots, n \quad (12)$$

where $y_i$ indicates the class to which the correspondent point $x_i$ belongs, and $p$ is the number of features describing
the vectors $x$. Any hyper-plane can be expressed in the following form:

$$w \cdot x - b = 0$$  \hfill (13)$$

that is the dot product between the normal vector ($w$) to the hyper-plane and a vector $x$. To separate linearly separable data, two hyper-planes can be identified. The corresponding region between them, where no data points are, is usually known as margin. The two hyper-planes are usually defined as:

$$w \cdot x - b = 1$$ \hfill (14)$$

and

$$w \cdot x - b = -1$$ \hfill (15)$$

with $2/\|w\|$ the size of the margin (i.e. distance between the two hyper-planes).

To avoid data points falling within the margin, the following constraint has to be defined:

$$y_i(w \cdot x_i - b) \geq 1 \quad i = 1, \ldots, n$$ \hfill (16)$$

Therefore, the hard margin formulation consists of minimizing $\frac{1}{2} \|w\|^2$ subject to the previous constraint.

Figure 3 presents an example of SVM classification for linearly separable data. However, hard margin SVM is effective only when data is linearly separable, a situation quite rare in real world problems. Thus, the soft margin SVM has been proposed (Scholkopf & Smola 2002): it relaxes separation constraints in order to admit some classification errors which are limited through a penalization term in the objective function. The C-SVM classification is an implementation of the soft margin where $C$ is a regularization parameter for setting the trade-off between minimization of classification error and maximization of margin.
In C-SVM, the constraint has to be modified as follows:

\[ y_i(w \cdot x_i - b) \geq 1 - \xi_i \quad i = 1, \ldots, n \quad \text{and} \quad \xi_i \geq 0 \quad (17) \]

while the objective function becomes:

\[ \min_{w, b, \xi} \left\{ \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} \xi_i \right\} \quad (18) \]

Nevertheless, both hard and soft margins work with a linear hyper-plane, a too strong precondition in real-world classification problems. The \textit{kernel trick} has been proposed (Scholkopf & Smola 2002) in order to perform – implicitly – a mapping of the instances from the original space (\textit{input space}) in a new one (\textit{feature space}) in which they could be hopefully linearly separated. In Figure 4, an example of this mapping is depicted.

Several types of kernel have been proposed (e.g. polynomial, radial basis functions, sigmoid), each one with at least an internal parameter to be set up for mapping (Scholkopf & Smola 2002).

\section*{Sensor location}

Both spectral clustering and SVM classification use variations in pressure and flow at the monitoring points as input data. Thus, sensor location logically affects the performance of the two analytical processes. The greater the number of sensors deployed, the higher is the capability to identify clusters with high quality as well as a reliable SVM classifier. However, complete deployment implies high costs for equipment and installation as well as useless redundancy of information.

To address cost-effective sensor placement, we propose to apply clustering to the leakage scenarios dataset but with another logic: clustering is applied to features which are variations in pressure and flow at each junction and each pipe of the WDN, respectively, and separately for the two types of meter. As a result, all the junctions and pipes, separately, showing similar variations in pressure and flow according to the several simulated leaks, are grouped...
together. Only the centroids (i.e. the most representative junctions and pipes) of the clusters are selected as the most relevant monitoring points, which are pressure meters in the case of junctions and flow meters in the case of pipes.

The number of clusters, corresponding to the number of pressure and flow meters to be deployed, affects both the quality of leakage localization (in particular LI, QL and LI⁰) as well as costs. To identify the best trade-off between these two aspects, a scatter plot has been drawn for each index (Figures 5–7). Costs for sensors have been set to 1 for each pressure meter and to 10 for each flow meter; the possible configurations for sensor deployment are related to 7, 10 or 13 pressure meters and 1, 2 or 3 flow meters.

Planning a sensor placement of seven pressure and three flow meters appears to be, globally (LI⁰), the best choice in terms of trade-off between leakage localization and deployment costs. In particular, further increasing the number of pressure meters to 10 improves LI while decreasing QL; on the other hand, no further improvement in LI is obtained by increasing to 13 the number of pressure meters, while QL does not change significantly.

Figure 8 depicts the identified sensors placement; flow meters are indicated with numbers and pressure meters with letters.

RESULTS

The results are presented, according to the best sensor placement identified as described above. The overall number of leakage scenarios that have been generated is 3,150, obtained by placing a leak, in turn, in each pipe, and varying its severity between 10 different values (Figures 9–14).

Leakage scenario clustering

As first result, LI, QL and LI⁰ obtained through spectral clustering are reported, depending on:
- number of scenarios clusters (K), from 3 to 22;
- number of relevant eigen-vectors (l), 3 or 4;
- type of clustering algorithm applied in the eigen-space: farthest first (S-FF) and simple K-means (S-SKM).

Taking into account the definition of LI, it is quite easy to understand that increasing K improves the LI while it reduces the QL. The figures show that spectral clustering internally using S-SKM performs better than the one using S-FF and is also not so dependent on the number of relevant eigen-vectors (l), in particular according to LI.

![Figure 5](https://iwa.silverchair.com/aqua/article-pdf/64/5/567/399844/jws0640567.pdf) | Costs for sensors (y-axis) versus LI (x-axis). Labels are PnFm, with n number of pressure meters and m number of flow meters. Cost of a pressure meter is set to 1 and cost of a flow meter is set to 10.
The best configuration selected in this study is the algorithm S-SKM with \( l = 3 \) and \( K = 12 \). Although the highest value of \( L_{I^*} \) is not reached in correspondence to \( K = 12 \) (i.e. with respect to \( l = 3 \), the maximum of \( L_{I^*} \) is in \( K = 10 \)), we decided to lose something in terms of QL to gain something in terms of localization, still keeping restricted the number of clusters. This decision also takes into account the results provided by the not selected algorithm (S-FF).

To demonstrate the benefits provided by spectral clustering with respect to traditional clustering algorithms, the values of LI, QL and \( L_{I^*} \) are reported, in three different figures, depending on:
number of clusters ($K$);
- traditional clustering algorithm (in the input space and not in the eigen-space): FF and SKM.

Results are compared to the best spectral clustering configuration $S$-$SKM^*$ (i.e. $K = 12$, $l = 3$, algorithm $S$-$SKM$), and its relation with $K$. 

**Figure 8** | Best sensors placement identified.

**Figure 9** | $L_I$ for spectral clustering, depending on number of clusters ($K$), number of relevant eigen-vectors ($l$) and type of clustering algorithm applied in the eigen-space.
Performances provided by spectral clustering are clearly better than those offered by clustering algorithms which are not graph-based. Finally, Figures 15 and 16 show the best and the worst clusters in terms of LI*.

Support vector machines: localizing leaky pipes effectively

Several configurations of the SVM classification have been evaluated to identify the most reliable relationship
between variations in pressure and flow and the possible location of the leak. The evaluation has been performed via 10-fold cross-validation, a technique that uses the entire dataset to learn the relationship and then test it, giving an estimation of the reliability in predicting the cluster associated with new vectors of hydraulic variations. Accuracy measure has been adopted as the performance measure, i.e. the number of vectors of hydraulic variations correctly associated to the cluster provided by the spectral clustering process.
The best SVM classifier was shown to be a C-SVM classifier, with \( C = 1 \) and a radial basis function kernel (with internal kernel parameter \( \gamma = 4.0 \)).

This classifier has been further tested in an independent test set, related to a new leakage scenario generation process performed by using values of severity different from those already adopted (i.e. new leaks). In this case, spectral clustering has not been applied to this test set; the trained SVM classifier provides an estimation of the cluster that spectral clustering should assign to each new vector of variations. If the pipe associated with a variation appears in the set of distinct pipes associated with the predicted scenarios cluster, this is counted as a success in localization.

The reliability of the model is computed as the number of successes with respect to the sum of the number of correct and incorrect localizations. This index has been
computed for each cluster, in order to evaluate whether reliability in localization may differ in clusters. Figure 17 compares the value obtained from the training and test sets, showing that localization performances are highly retained from training to test and proving the approach is highly reliable.

DISCUSSION

This study proposed further improvements to the analytical leakage localization approach already proposed by the authors and aimed at improving leakage management in urban WDNs. The approach adopts: simulation of several leaks (varying their location and severity), graph-based clustering algorithm (spectral clustering) of the obtained pressure and flow variations and, finally, classification learning (SVM) to identify a reliable relationship between variations in pressure and flow and leak location.

Owing to the inapplicability of the traditional clustering fitness measures, ad hoc indexes have been proposed to measure (i) the capability of the approach to localize a leak on a restricted set of pipes (LI) and (ii) the capability to avoid associating the same pipe with different types of hydraulic variations (QL). Finally, a combination of the two indexes ($LI^* = LI \times QL$) has been adopted to compare different configurations of the overall approach and select the best one.

As a further result, we propose an approach to support cost-effective sensor placement and used the best setting identified to test and evaluate their overall approach on a real case study, the Neptun DMA in Timisoara, Romania, one of the two pilots of the European project ICeWater.

A relevant result of this study, in particular with respect to our previous studies, is related to the adoption of SVM classification to identify, starting from the results provided by spectral clustering, a reliable relationship from variations in pressure and flow toward the leak location (i.e. a restricted set of pipes to check physically).

This new version of the approach offers several benefits; first of all, regression of the leak severity (Candelieri et al. 2013a, b) is no longer needed because leaks simulated in a pipe are put in the same cluster irrespective of their severity (as proved by the QL index). As a further benefit, the application of SVM classification permits reduction of computational costs related to spectral clustering: a smaller – even if significant – set of leakage scenarios is required to perform spectral clustering while SVM will approximate the non-linear mapping from the input space (variations...
in pressure and flow at the monitoring points) to the eigenspace spanned by the most relevant eigen-vectors of the Laplacian matrix.

While the proposed approach aims at improving leakage management through a more accurate and cost-effective analytical leakage localization solution, it also provides effective strategies for reducing computational costs related to the application of graph-based analysis on large dataset (e.g. generated through extended simulation).

The main limitations of this study are related to the validation of the approach only in a simulated environment and taking into account only a DMA. Further investigations, with respect to these issues, are going to be performed in the following activities of the ICeWater project. In particular, a sensitivity analysis has already been performed and relative results have been submitted for a forthcoming publication, taking into account different sources of ‘noise’: errors on the pressure and flow measurements, errors in modeling demand at the junctions and errors in modeling the WDN hydraulic infrastructure. On the other hand, the approach will be also applied and validated in a wider WDN, in particular the Pressure Management Zone in Milan, the Italian pilot of ICeWater, namely Abbiategrasso. Furthermore, a ‘real world’ validation is also planned for the final year of the ICeWater project, consisting of opening some hydrants to create controlled/artificial leaks and evaluate the rate of successful localizations provided by the approach developed.

These further validation activities will allow for a more comprehensive evaluation and comparison with other recent leakage localization approaches (Farley et al. 2010a, b, 2013; Mounce et al. 2010; Wu et al. 2010; Pérez et al. 2011; Preis et al. 2011), in order to evaluate the real benefits provided by spectral clustering in capturing, more accurately, the non-linear relation between the cause (location and severity of a leak) and the most probable corresponding effect (variations of pressure and flow at relevant monitoring points). Finally, one of the advantages of the proposed approach is to link the optimal sensor placement directly to the localization performances – even if computed in a simulation-based environment – enabling a decision-making process aimed at identifying the best trade-off between accuracy in leakage localization and budgetary constraints.

ACKNOWLEDGEMENT

This work has been partially supported by the European Union ICeWater project – FP7-ICT 317624 (www.icewater-project.eu).

REFERENCES

Izquierdo, J., Herrera, M., Montalvo, I. & Pérez-García, R. 2011 Division of water supply systems into district metered areas...

Jaakkola, T. 2006 Course Materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology.


Scholkopf, B. & Smola, A. J. 2002 Learning with Kernels. Support Vector Machines, Regularization, Optimization and Beyond. Massachusetts Institute of Technology, USA.


First received 20 March 2014; accepted in revised form 1 April 2015. Available online 4 May 2015.