Journal of Water Supply: Research and Technology - AQUA

Cost-Effective Sensors Placement and Leak Localization - Neptun pilot of the ICeWater project

--Manuscript Draft--

Manuscript Number:	
Article Type:	Editorial Office Upload
Section/Category:	CCWI2013
Corresponding Author:	Antonio Candelieri Consorzio Milano Ricerche Milano, ITALY
First Author:	Antonio Candelieri
Order of Authors:	Antonio Candelieri
Abstract:	This papers extends the research activities performed by the authors in developing and applying an approach to analytically identify leaks within a Water Distribution Network (WDN), by combining hydraulic simulation (EPANET) and Network Science based Data Analysis techniques. The software model of the WDN is used to run several "leakage scenarios", by varying leak location (pipe) and severity, and to build a dataset with the corresponding variations in pressure, at nodes, and flow, on pipes, induced by the leak. All junctions and pipes are considered for potential pressure and flow sensors deployment; a clustering procedure on these potential locations is then performed to identify the most relevant nodes and pipes, while costs for pressure and flow meters are considered to select the combination which guarantees the best trade-off between reliability in localizing leaks and overall cost. 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 is adopted to group together similar scenarios in the eigen-space spanned by the most relevant eigenvectors of the Laplacian Matrix of the graph. This approach proved to

Cost-Effective Sensors Placement and Leak Localization – Neptun pilot of the ICeWater project

Antonio Candelieri[a*], Davide Soldi[b], Francesco Archetti[a,b]

[a]*Consorzio Milano Ricerche, via Roberto Cozzi 53, Milan, 20126, Italy*
[b]*Department of Computer Science, Systems and Communication, University of Milano Bicocca, viale Sarca 336, Milan, 20126, Italy*

Abstract
This papers extends the research activities performed by the authors in developing and applying an approach to analytically identify leaks within a Water Distribution Network (WDN), by combining hydraulic simulation (EPANET) and Network Science based Data Analysis techniques. The software model of the WDN is used to run several "leakage scenarios", by varying leak location (pipe) and severity, and to build a dataset with the corresponding variations in pressure, at nodes, and flow, on pipes, induced by the leak. All junctions and pipes are considered for potential pressure and flow sensors deployment; a clustering procedure on these potential locations is then performed to identify the most relevant nodes and pipes, while costs for pressure and flow meters are considered to select the combination which guarantees the best trade-off between reliability in localizing leaks and overall cost.
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 is adopted to group together similar scenarios in the eigen-space spanned by the most relevant eigenvectors of the Laplacian Matrix of the graph. This approach proved to be more effective than other traditional techniques which work directly in the space of pressure and flow variations. 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.

Keywords: Water Distribution Networks; Leakage Management; Leakage Localization; Spectral Clustering; Support Vector Machines

*Corresponding author: Antonio Candelieri, candelieri@milanoricerche.it, +39 02 6448 2185

1. Introduction

Nowadays an innovative and more smart management of urban water distribution networks (WDN) is needed to achieve higher levels of efficiency, exploiting the huge amount of data already generated by and stored into the ICT systems adopted in WDN management operations, such as Supervisory and Control Data Acquisition (SCADA) systems and hydraulic simulation, as well as new data made available by the introduction of smart metering solutions (Automatic Metering Reader, AMR). The availability of data, the exploitation of the simulation software and the recent research on data analysis enable the shift toward a smart water management paradigm.

As already reported in (Alegre et al., 2006) the International Water Association (IWA) highlighted the relevance to improve the leakage management process also providing some specific performance indicators. In (Puust, 2010) a formalization of the leakage management process is proposed, 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 (NRW), increasing costs for energy and rehabilitation, in spite of more 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, probabilistic modeling have been investigated (Poulakis et al. 2003, Caputo and Palagge 2003, Xia et al. 2006, Sivapragasam et al. 2007, Behzadian et al. 2009, Xia and Guo-Jin 2010, Lijuan et al. 2012, Nasir et al. 2012).

While most of these approaches are focused on localizing a leak on pipes, another recent and relevant research work proposes a combination between hydraulic simulation and Support Vector Machines (SVM) classification to identify leaks on junctions according to the pressure and flow values (Mashford et al. 2012).

With respect to machine learning, another relevant research filed to support a more effective and efficient leakage management is focused on leaks and bursts detection through the analysis of real-time sensors data, without using any simulation (Romano et al., 2011). 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 presents further research activities performed by the authors in the definition and application of a reliable approach for localizing leaks, analytically, going beyond the results already presented in their previous works (Candelieri and Messina 2012, Candelieri et al. 2013a, Candelieri et al. 2013b).

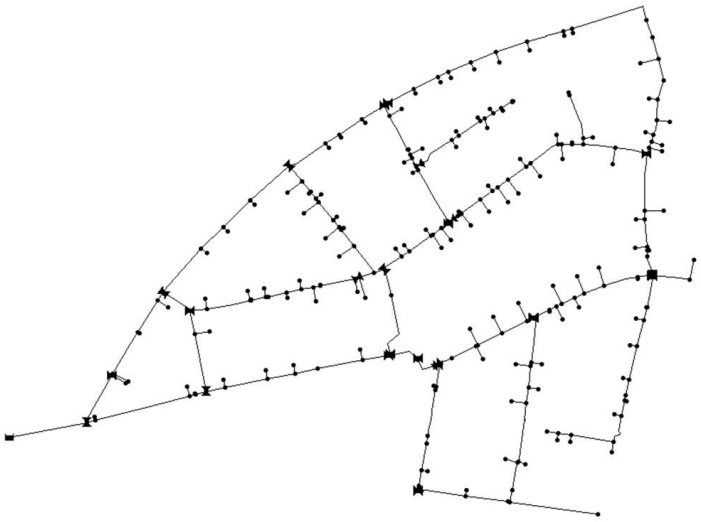
More in detail, contribution of this paper concerns different topics: *i)* a deeper investigation of the benefits provided by Network Science based machine learning approaches (i.e. Spectral Clustering) with respect to more "traditional" ones *ii)* the proposition of a method to support a cost-effective placement of flow and pressure meters, with respect to reliability of analytical leakage localization and *iii)* the identification, through Support Vector Machines(SVM), 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 related to a real test case, a District Meter Area (DMA) of the WDN in Timisoara, Romania, one of the two pilots of the FP7-ICT project ICeWater co-funded by the European Commission.

The rest of the paper is organized as follows: section 2 describes the pilot, the hydraulic simulation process, the methodological background on Spectral Clustering; section 3 describes the adoption of SVM classification; in section 4 the sensor location approach is proposed; section 5 reports the experimental results. A discussion about the perspective of the approach is finally provided.

2. Materials and Methods

Description of the pilot



In the following Figure 1 the ICeWater pilot (Neptune) in Timisoara, Romania, in depicted. It is a DMA whose pipe infrastructure is long about 4000m, overall. Pipes length ranges from 0.034m to 83.808m (12.778m ± 13.156m), pipes diameter ranges from 50mm to 500mm (106.917mm ± 76.450mm) and roughness varies from 1 to 150 (111.478 ± 16.084).

Fig. 1. Neptun: the pilot DMA of the ICeWater project, in Timisoara.

Hydraulic Simulation of Leakage Scenarios through EPANET

The hydraulic simulation software EPANET, free downloadable from the Environmental Protection Agency web site (http://www.epa.gov/nrmrl/wswrd/dw/epanet.html) is a widely used tool for modeling WDNs, performing *what-if* scenarios 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 in placing, in turn, a leak on each pipe and varying its severity in a given range. 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 firstly described in (Candelieri and Messina 2012).

Clustering Leakage Scenarios and Quality Measure

Clustering leakage scenarios previously generated through EPANET is 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 pipe 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 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 leak has required the definition of a specific measure, namely "Localization Index", already proposed in a previous work of the authors. The Localization Index for each cluster (LI_k) requires to retrieve the information on leaky pipe 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:

$$LI_k = \frac{|pipes| - |pipes_k|}{|pipes| - 1}$$

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 of LI_k is $LI_k = 1$ that is obtained when the cluster k contains scenarios all related to leaks simulated only on one pipe (i.e., $|pipes_k| = 1$). On the other hand, the minimum value of LI_k is $LI_k = 0$ that is obtained if the cluster k contains scenarios referred to all the pipes of the WDN (i.e., $|pipes_k| = |pipes|$).

While in the previous work of the authors the overall localization index (LI) has been computed as the simple average of LI_k , in this case the average has been weighted by the number of distinct pipes in each cluster:

$$LI = \frac{\sum_{k=1}^{K} LI_k |pipes_k|}{\sum_{k=1}^{K} |pipes_k|}$$

where K is the overall number of cluster.

In this work another relevant measure is proposed to evaluate how much a clustering algorithm is able to put in the same cluster scenarios related to same (leaky) pipe and to all the different severity values. This index has been named "Quality of Localization" and is defined, for each cluster k, as:

$$QL_k = \frac{\sum_{p_j \in S} \frac{n_j^k}{|S|}}{n_p^k}$$

where S is the set of different severity used (and $|S|$ is the overall number of severity values used), p_j is a pipe having a leak whose severity is the j^{th} in S, n_j^k is the number of scenarios in cluster k associated to 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 is $\underline{QL_k} = 1$ that is obtained when in the cluster k are all the scenarios related to all the severity values of the correspondent leaky pipes.

The overall Quality of Localization for a clustering algorithms is given by the average of QL_k.

Finally, a global index LI* is defined, combining Localization Index and Quality of Localization

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

Spectral Clustering

Although Spectral Clustering (Luxburg 2007, Jaakkola 2006) 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 on not-relational data points datasets.

Final goal is the same both for traditional and spectral clustering, that is 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:

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

A n-dimensional vector p (i.e., 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 :

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

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

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

where L_{ij} are the entries of the Laplacian matrix, the core of spectral clustering. Different alternative definitions have been proposed and studied through graph theory (Chung, 1997); the usually adopted 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:

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

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

The most important properties of the L matrix are:

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

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

$$L_{norm} = I - D^{-1/2} A D^{-1/2}$$

The combinatorial complexity of the minimizing (4) can be prohibitive for real world networks. 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 2nd smallest eigenvector 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 to implement recursive bi-partitioning spectral clustering approaches (Hagen and 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 eigenvectors (Luxburg, 2007; Ng et al., 2001). The relevant eigenvectors are the first l smallest: the l-th eigenvalue is the one showing a sufficiently large variation in the *eigengap*, that is the difference between two successive eigenvalues in the list of eigenvalues sorted in ascending order.

For example, the K-partitioning approach proposed in (Shi and Malik, 2000), consists in 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 (eigenvectors corresponding to the l smallest eigenvalues). Any other traditional clustering algorithm may be applied in the eigenspace.

3. Identifying leaky pipes through Support Vector Machines

After clustering the leakage scenarios, the next step consists in discovering a reliable relation between the variations in pressure and flow, due to a leak, and the correspondent cluster, which permits to retrieve the set of correspondent leaky pipes. This relation allows reducing time for investigations and rehabilitation: when a leak is detected (e.g., with traditional methods, such as Minimum Night Flow analysis (Liemberger and Farley, 2004, Behzedian et al. 2009, and 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 physically check.

In their previous study, the authors proposed to compare the obtained vector of values with those related to the centroids of the clusters, and selected the cluster associated to the most similar centroid. However, since the Spectral Clustering procedure implicitly applies a non linear transformation from the *Input Space* (related 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.

In order to improve the reliability of the localization process, a Support Vector Machine (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 vectors of variations in pressure and flow belongs to.

The following Figure 2 shows the overall workflow, presenting the mapping performed by the Spectral Clustering. The SVM permits to apply 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 to adopt a reliable approximation of Spectral Clustering to identify the correspondent scenarios cluster for any new vector of variations in flow and pressure.

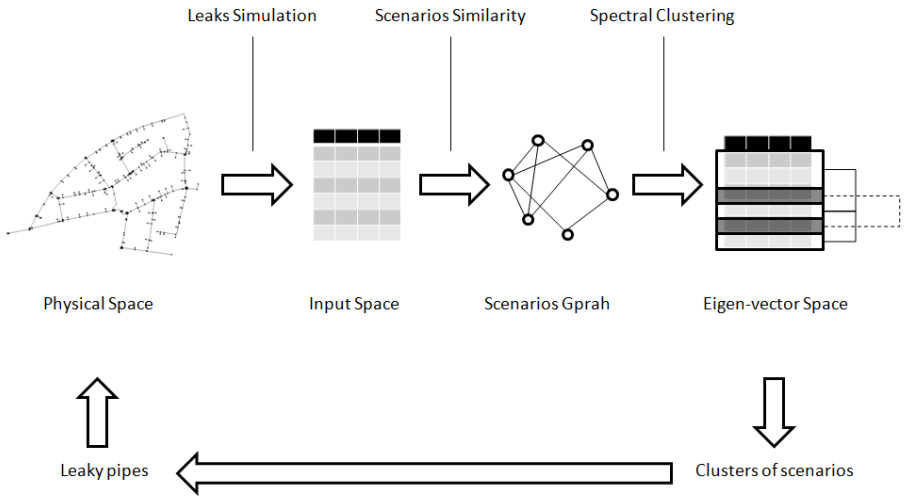


Fig. 2. Overall approach proposed: Spectral versus traditional clustering of leakage scenarios.

Support Vector Machines classification

The basic idea of SVM (Vapnik 1998) consists in searching for a hyper-plane to optimally separate instances (represented as vectors) belonging to different classes and, contemporary, 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:

$$D = \{(x_i, y_i) \mid x_i \in \Re^p, y_i \in \{-1,1\}\} \text{ with i} = 1, \ldots, \text{n}$$

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 following form:

$$w \cdot x - b = 0$$

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

$$w \cdot x - b = 1$$

and

$$w \cdot x - b = -1$$

with $2/||w||$ the size of the margin (i.e., distance between the two hyper-planes).
In order to avoid data points falling within the margin, the following constraint has to be defined:

$$y_i(w \cdot x_i - b) \geq 1 \quad \text{i} = 1, \ldots, n$$

Therefore, the hard margin formulation consists in minimizing $\frac{1}{2} ||w||^2$ subject to the previous constraint.
In the following Figure 3 an example of SVM classification for linearly separable data is presented.

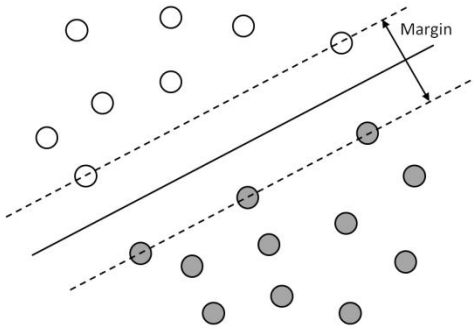


Fig. 3. An example of linear separation through margin maximization

However, hard margin SVM is effective only when data is linearly separable, a situation quite rare in real world problem. Thus, the *soft margin SVM* has been proposed: 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 modified as follows:

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i \quad i = 1, \ldots, n \quad \textit{and} \quad \xi_i \geq 0$$

while the objective function becomes:

$$\min_{w,\xi,b} \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \right\}$$

Nevertheless, both hard and soft margin works with a linear hyper-plane, a too strong precondition in real world classification problems. The *kernel trick* has been proposed in order to perform - implicitly - a mapping of the instances form the original space (*Input Space*) in a new one (*Feature Space*) in which they could be hopefully linearly separated. In Figure 4 an example of this mapping is depicted.

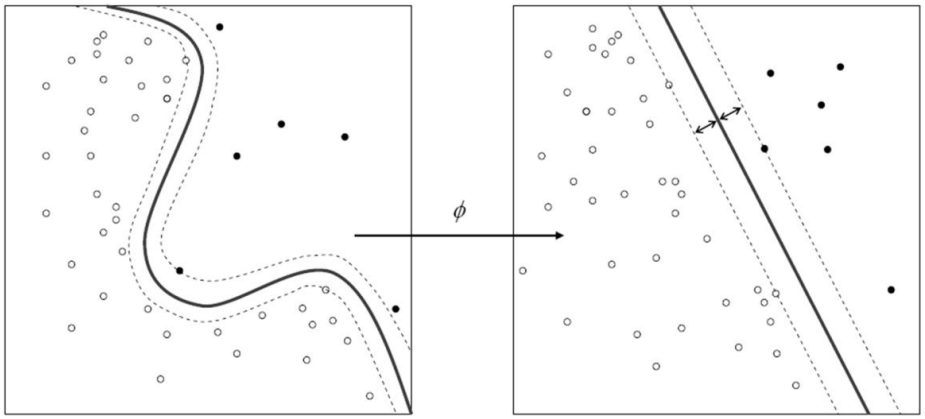


Fig.4. From Input to Feature Space through kernel trick: the linear separation in the Feature Space (right) corresponds to a non-linear separation in the original Input Space (left)

Several types of kernel have been proposed (e.g. Polynomial, Radial Basis Functions, Sigmoid, etc.) each one with at least an internal parameter to be set up for mapping (Scholkopf and Smola 2002).

4. Sensors Location

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

To address cost-effective sensor placement, the authors propose to apply, again, clustering on the leakage scenarios dataset but with another logic: clustering is applied on features, that are variations in pressure and flow at each junction and each pipe of the WDN, respectively, and separately for the two types of meters. As 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, that are a pressure meters in the case of junctions and a 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 Li*) as well as costs. In order to identify the best trade-off between these two aspects, a scatter plot has been drawn for each index (Figures 5, 6 and 7). Costs for sensors have been set 1 for each pressure meter and 10 for each flow meter; the possible configurations for sensors deployment are related to 7, 10 or 13 pressure meters and 1, 2 or 3 flow meters

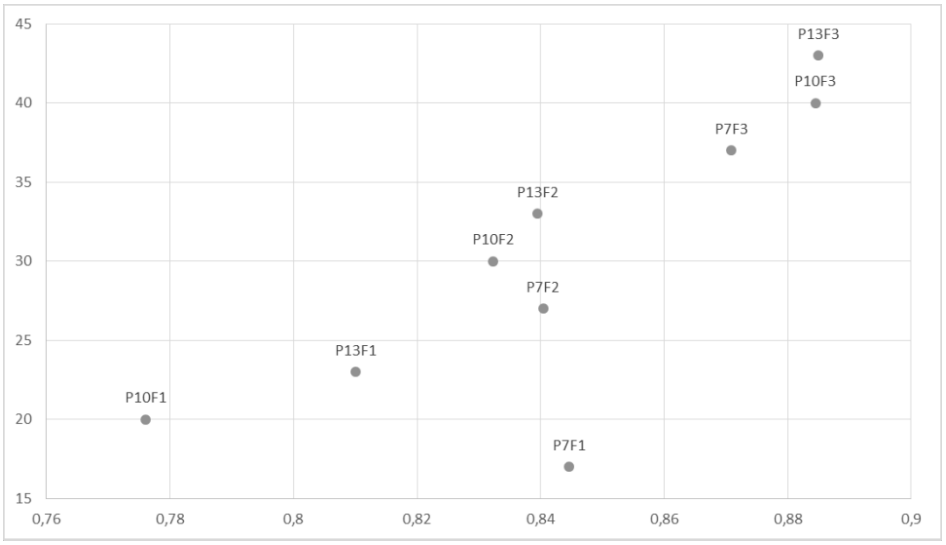


Fig. 5. Costs for sensors (y-axis) versus LI (x-axis). Labels are P*n*F*m*, with *n* number of pressure meters and *m* number of flow meters. Cost of a pressure meter is set to 1 and cost of a flow meter is et to 10.

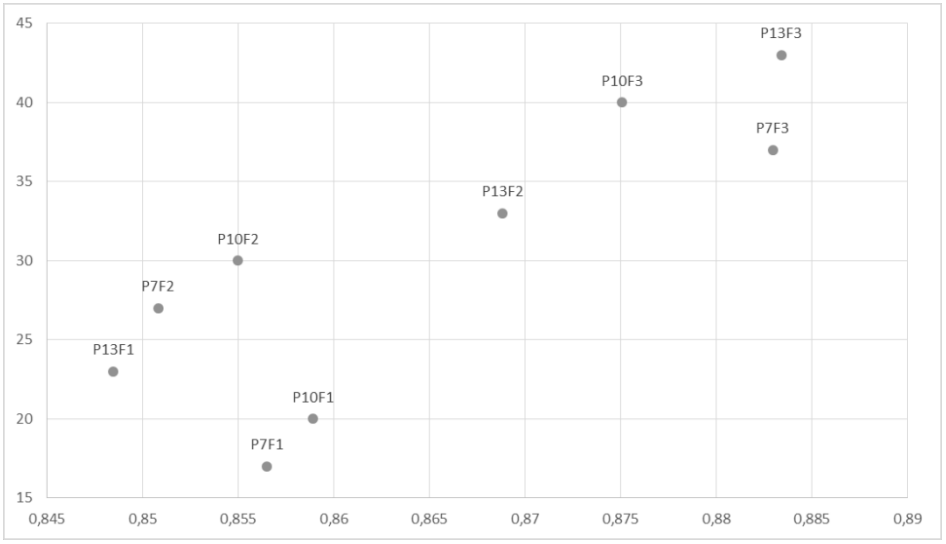


Fig. 6. Costs for sensors (y-axis) versus QL (x-axis).

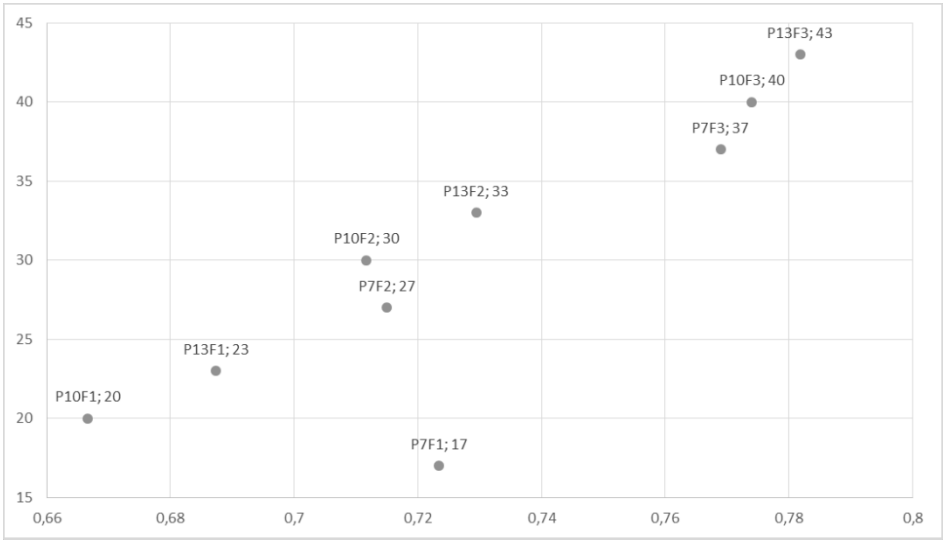


Fig. 7. Costs for sensors (y-axis) versus LI* (x-axis).

Planning a sensor placement of 7 pressure and 3 flow meters appears to be, globally (LI*), 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.

In the following Figure 8, the identified sensors placement is depicted, where flow meters are indicated with numbers and pressure meters with letters.

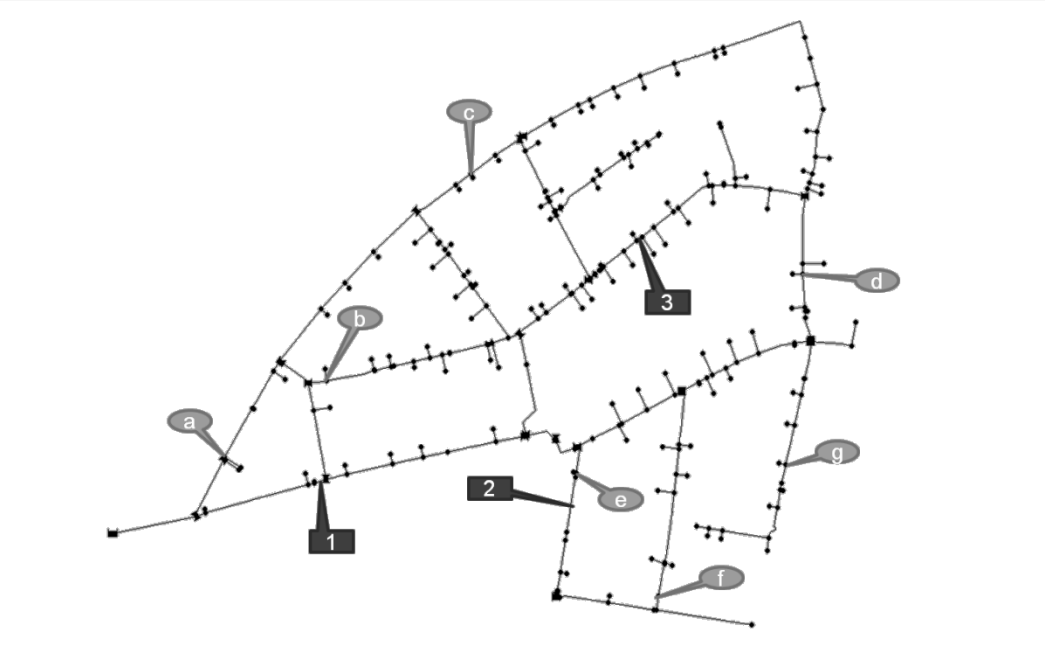


Fig. 8. Best sensors placement identified.

5. Experimental Results

In this section, the results are presented, according to the best sensors placement identified and presented in the previous section.

The overall number of leakage scenarios that have been generated is 3150, obtained by placing a leak, in turn, on each pipe, and varying its severity among 10 different values.

Results on Leakage Scenarions Clustering

As first result, LI, QL and LI* obtained through Spectral Clustering are reported depending on:

- number of scenarios clusters (*K*), form 3 to 22;
- number of relevant eigen-vectors (*l*), 2 or 3;
- type of clustering algorithm applied in the eigen-space: Farthest First (*S-FF*) and Simple *K*-means (*S-SKM*)

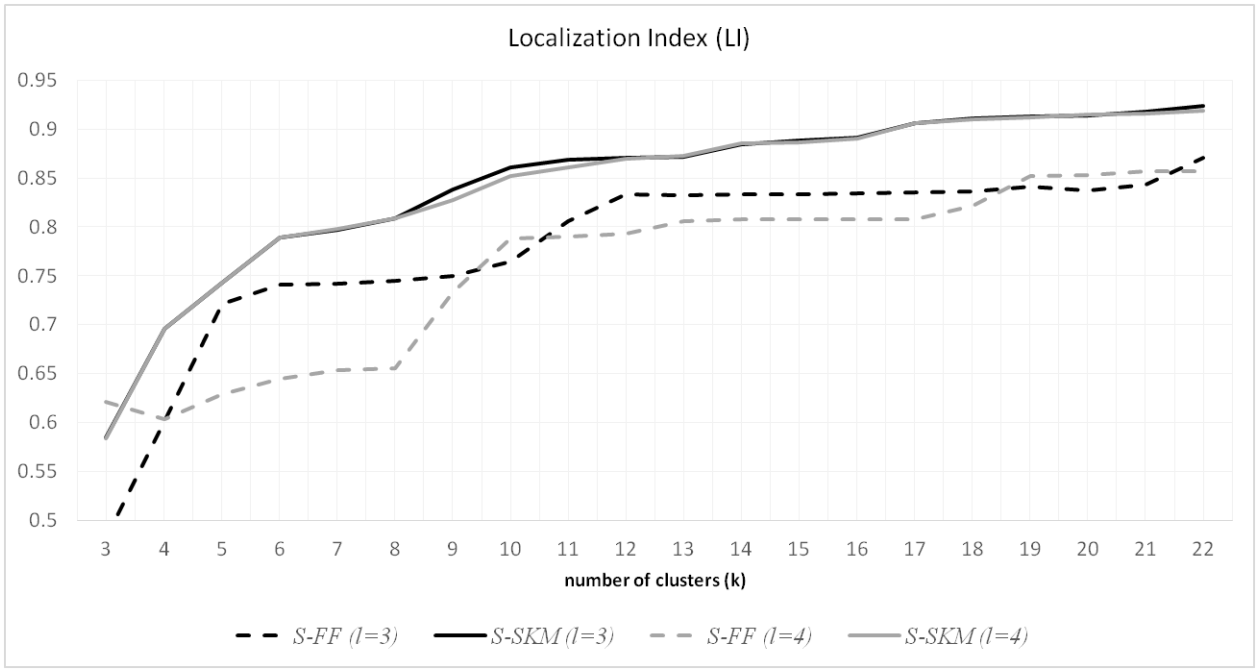


Fig. 9. Localization Index (LI) for Spectral Clustering, depending on number of clusters (*K*), number of relevant eigen-vectors (*l*) and type of clustering algorithm applied in the eigen-space.

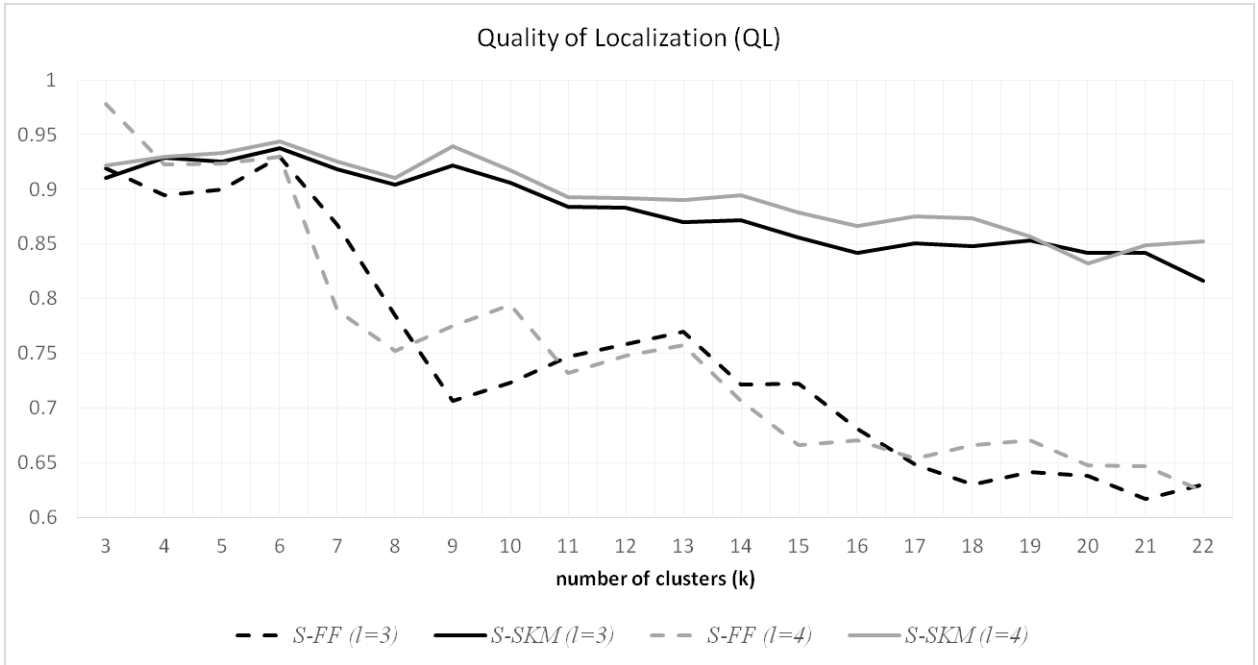


Fig. 10. Quality of Localization (QL) for Spectral Clustering, depending on number of clusters (*K*), number of relevant eigen-vectors (*l*) and type of clustering algorithm applied in the eigen-space.

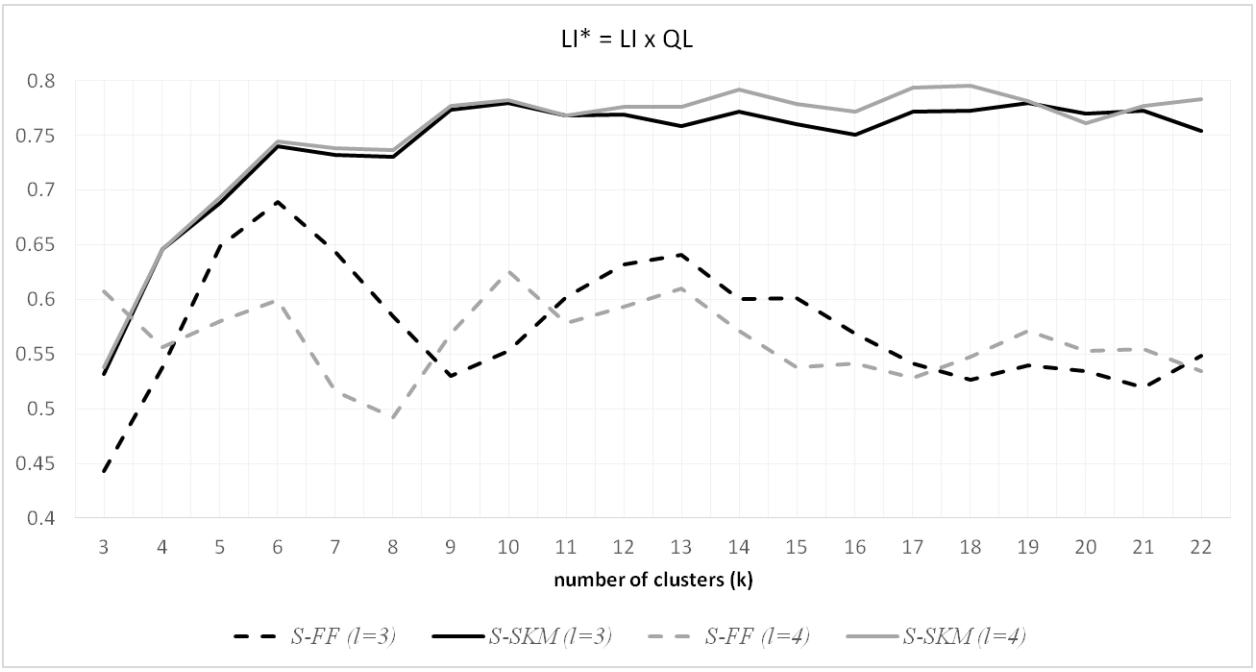


Fig. 11. LI* for Spectral Clustering, depending on number of clusters (*K*), number of relevant eigen-vectors (*l*) and type of clustering algorithm applied in the eigen-space.

Taking into account the definition of LI, it is quite easy to understand that increasing *K* improves the Localization Index (LI) while reduces the QL. Having a look at the figures it is also easy to note that Spectral Clustering internally using Simple *K*-means (*S-SKM*) is most performing than the one using Farthest First (*S-FF*) and is also not so dependent on the number of relevant eigen-vectors (*l*), in particular according to LI.

The best configuration selected in this paper is the algorithm *S-SKM* with *l*=3 and *K*=12. Although the highest value of LI* is not reached in correspondence to *K*=12 (i.e., with respect to *l*=3, the maximum of LI* is in *K*=10), authors decided to lose something in terms of QL to gain something in terms of localization, still keeping restricted the number of clusters. This decision takes also into account the results provided by the not selected algorithm (*S-FF*).

In order to demonstrate the benefits provided by Spectral Clustering with respect to traditional clustering algorithms, the values of LI, QL and LI* 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): Farthest First (*FF*) and Simple *K*-means (*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*.

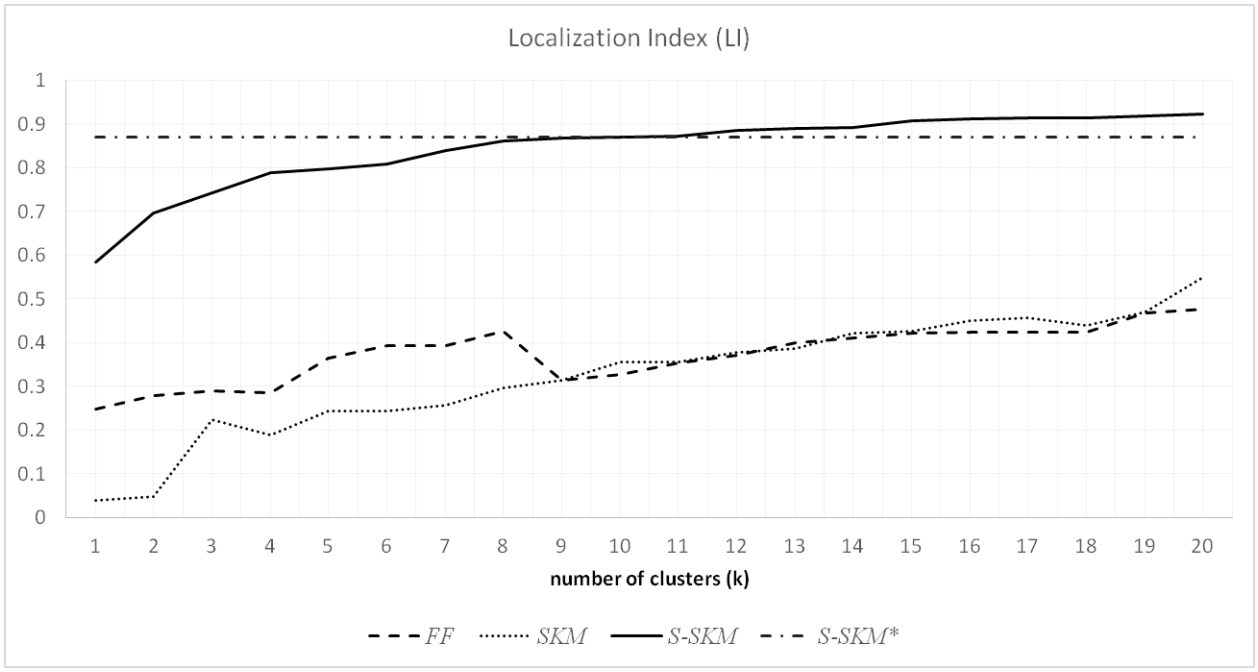


Fig. 12. Localization Index (LI):comparison between traditional clustering algorithms and best Spectral Clustering configuration, depending on number of clusters (*K*), number of relevant eigen-vectors (*l*) and type of clustering algorithm applied in the eigen-space.

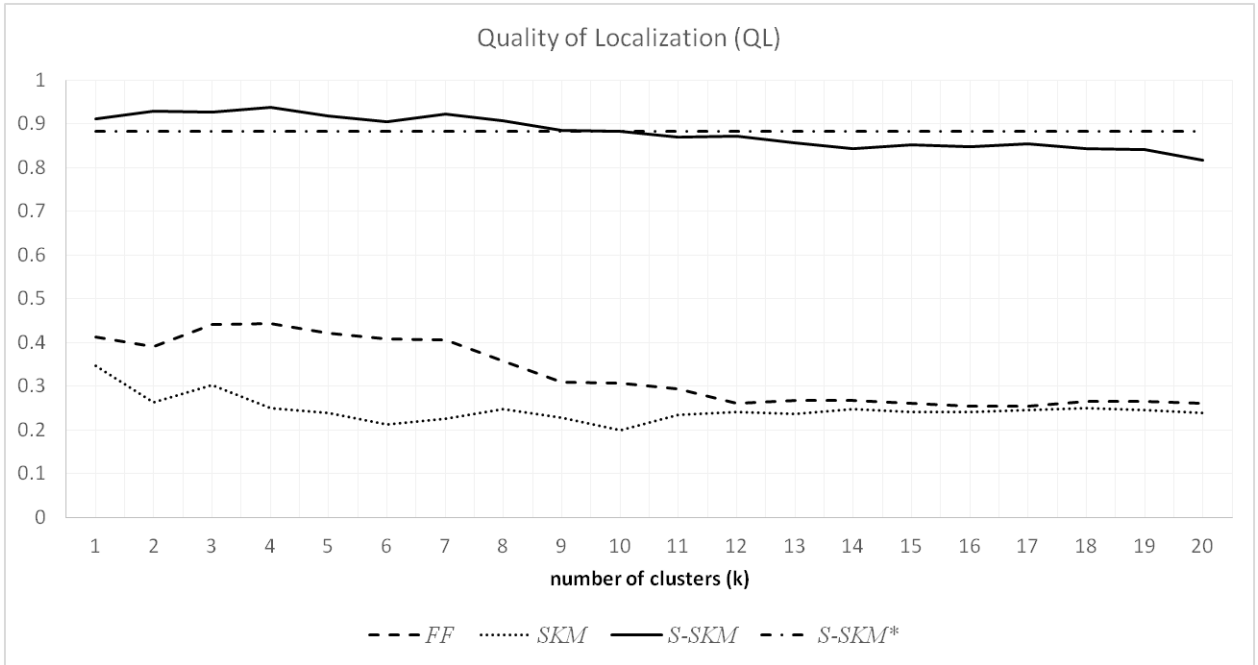


Fig. 13. Quality of Localization (QL):comparison between traditional clustering algorithms and best Spectral Clustering configuration, depending on number of clusters (*K*), number of relevant eigen-vectors (*l*) and type of clustering algorithm applied in the eigen-space.

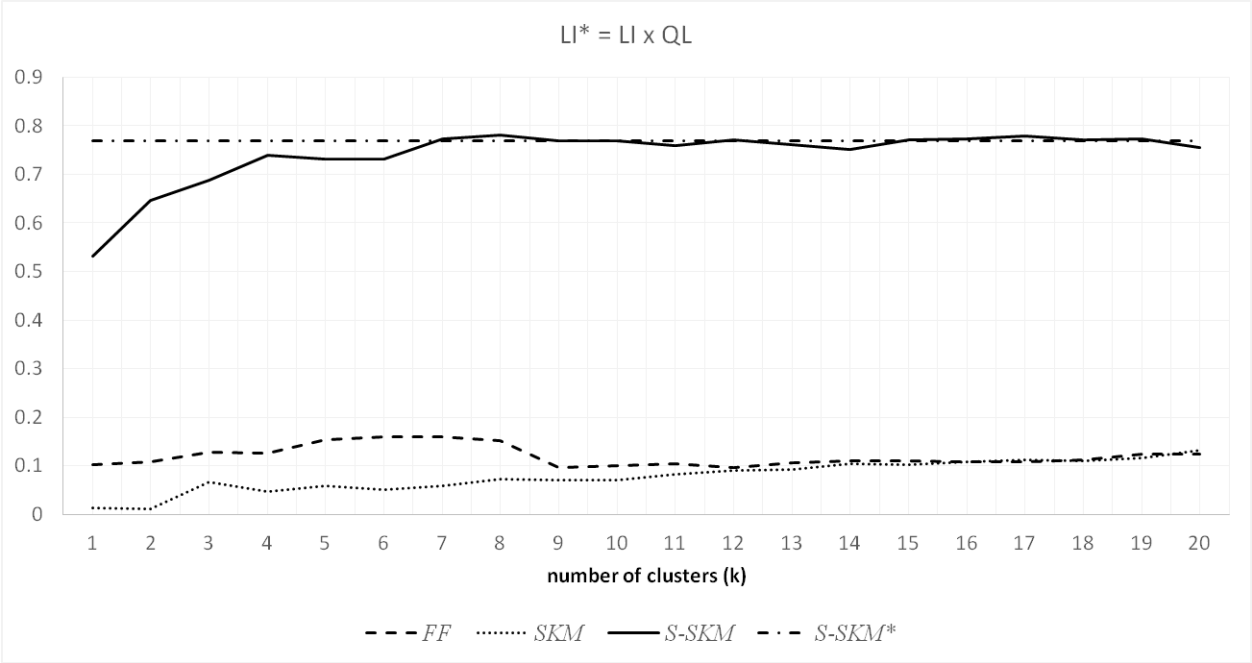


Fig. 14. LI*:comparison between traditional clustering algorithms and best Spectral Clustering configuration, depending on number of clusters (*K*), number of relevant eigen-vectors (*l*) and type of clustering algorithm applied in the eigen-space.

Performances provided by Spectral Clustering are clearly higher than those offered by clustering algorithms which are not graph-based. Finally, the following Figures 15 and 16 show the best and the worst clusters in terms of LI*.

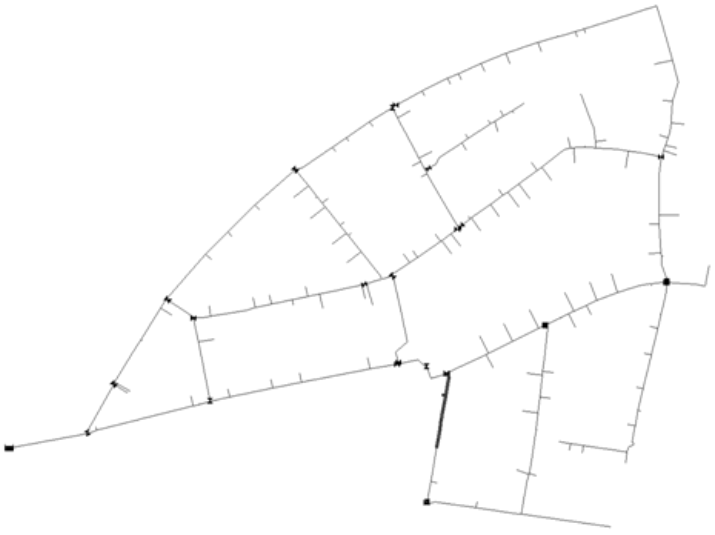


Fig. 15. Set of probably leaky pipes in one of the "best" clusters.

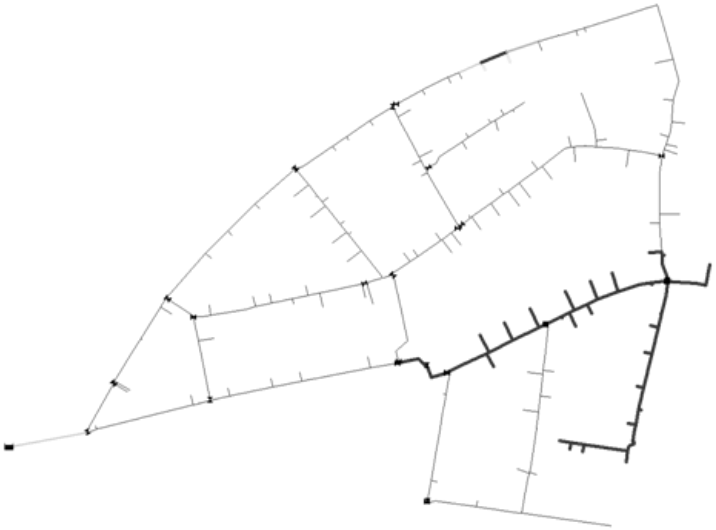


Fig. 16. Set of probably leaky pipes in one of the worst clusters.

Results on 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-folds 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 to new vectors of hydraulic variations. Accuracy measure has been adopted as performance measure, that is the number of vectors of hydraulic variations correctly associated to the cluster provided by the Spectral Clustering process.

The best SVM classifier resulted to be a C-SVM classifier, with C=1 and a Radial Basis Function (RBF) kernel (with internal kernel parameter $\gamma = 4.0$).

This classifier has been further tested on an independent test set, related to a new leakage scenarios generation process performed by using values of severity different from those already adopted (i.e. new leaks). In this case, Spectral Clustering has not be applied on 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 to a variation appears in the set of distinct pipes associated to 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 number of correct and incorrect localizations. This index has been computed for each cluster, in order to evaluate whether reliability in localization may differ on clusters. The following Figure 17 compares the value obtained on the training and test sets, showing that localization performances are highly kept from training to test and proving the approach is highly reliable.

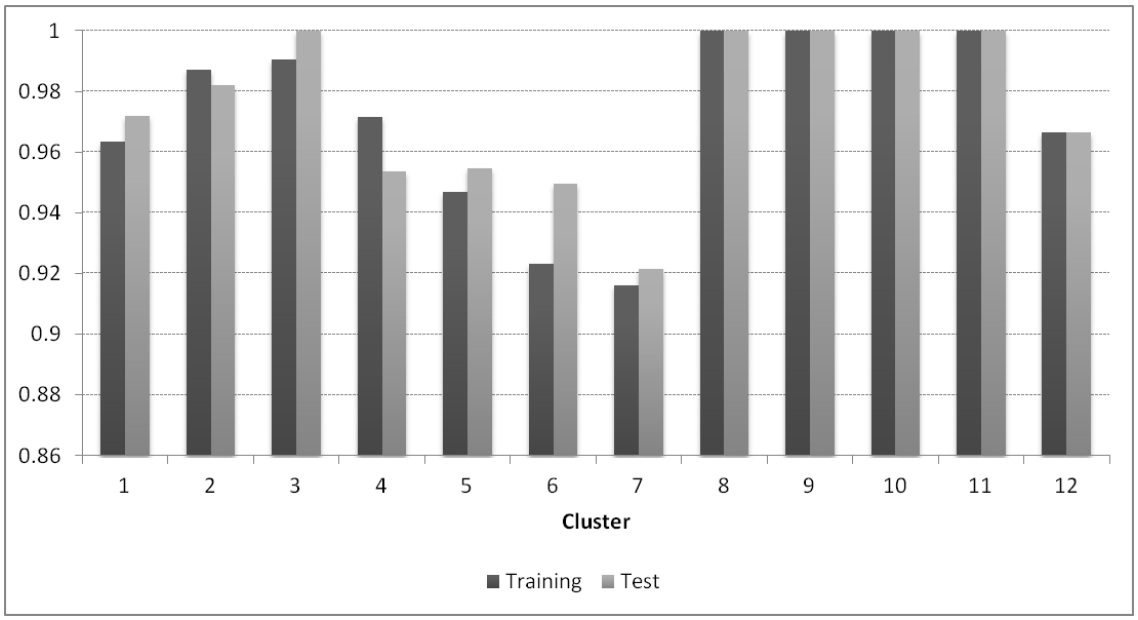


Fig. 17. Reliability of the localization, both on training set and independent test set.

Discussions

This study proposed further improvements to the analytical leakage localization approach already proposed by the authors and devoted to improve leakage management in urban WDN. 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 (Support Vector Machine) to identify a reliable relationship between variations, in pressure and flow, and leak location.

Since 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 (Localization Index, LI) and *ii)* the capability to avoid that the same pipe may be associated to different types of hydraulic variations (Quality of Localization, QL). Finally, a combination of the two indexes (LI* = LI x QL) has been adopted to compare different configurations of the overall approach and select the best one.

As further result, the authors have proposed 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 the previous works of the authors, is related to the adoption of Support Vector Machine classification to learn, staring form 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 physically check).

This new version of the approach offers several benefits; first of all, regression of the leak severity (Candelieri et al. 2013a,b) is not more needed because leaks simulated on a pipe are put in the same cluster irrespectively to their severity (as proved by the QL index). As a further benefit, the application of SVM classification permits to reduce 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 eigen-space spanned by the most relevant eigen-vectors of the Laplacian Matrix.

Finally, 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 data set (e.g., generated through extended simulation).

Acknowledgements

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

References

Alegre, H., Baptista, J.M., Cabrera, E., Cubillo, F., Duarte, P., Hirner, W., Merkel, W., Parena, R., 2006. Performance Indicators for Water Supply Services, Second Edition, IWA Publishing.

Behzadian, K., Kapelan, Z., Savic, D. A., Ardeshir, A., 2009. Stochastic sampling design using multi objective genetic algorithm and adaptive neural networks. Environmental Modeling and Software 24, 530–541.

Candelieri, A., Conti, D., Archetti, F., 2013a. A graph based analysis of leak localization in urban water networks, 12th International Conference on Computing and Control for the Water Industry, CCWI2013.

Candelieri, A., Archetti, F., Messina, E., 2013b. Improving leakage management in urban water distribution networks through data analytics and hydraulic simulation. WIT Transactions on Ecology and the Environment 171, 107-117.

Candelieri, A., Messina, E., 2012. Sectorization and analytical leaks localizations in the H2OLeak project: Clustering-based services for supporting water distribution networks management. Environmental Engineering and Management Journal 11(5), 953-962.

Caputo, A. C., and Pelagagge, P. M., 2003. Using Neural Networks to monitoring piping systems. Process Safety Progress 22(2), 119-127.

Chung, F., 1997. Spectral graph theory. Washington: Conference Board of the Mathematical Sciences.

Fiedler, M., 1973. Algebraic connectivity of graphs. Czechoslovak Mathematical Journal 23, 298–305.

Hagen, L. and Kahng, A., 1992. New spectral methods for ratio cut partitioning and clustering. IEEE Transacations on Computer-Aided Design 11(9), 1074-1085.

Izquierdo, J., Herrera, M., Montalvo, I., Pérez-García, R, 2011. Division of Water Supply Systems into District Metered Areas Using a Multi-agent Based Approach, In: Software and Data Technologies, Series Communications in Computer and Information Science, Cordeiro J., Ranchordas A., Shishkov B. (Eds.), Springer Berlin Heidelberg 50, 167-180.

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

Liemberger, R., Farley, M., 2004. Developing a nonrevenue water reduction strategy Part 1: Investigating and assessing water losses. In Proceeding of IWA WWC 2004 Conference, Marrakech, Morocco.

Lijuan, W., Hongwei, Z., and Hui, J., 2012. A Leak Detection Method Based on EPANET and Genetic Algorithm in Water Distribution Systems. Software Engineering and Knowledge Engineering: Theory and Practice – Advances in Intelligent and Soft Computing 14, 459-465.

Luxburg, U., 2007. A Tutorial on Spectral Clustering. Statistics and Computing 17(4), 1-32.

Mashford, J., De Silva, D., Burn, S., and Marney, D., 2012. Leak Detection in simulated water pipe networks using SVM. Applied Artificial Intelligence: An International Journal 26(5), 429-444.

Nasir, A., Soong, B. H., Ramachandran, S., 2010. Framework of WSN based human centric cyber physical in-pipe water monitoring system. 11th International Conference on Control, Automation, Robotics and Vision, 1257-1261.

Ng, A.Y., Jordan, M., Weiss, Y., 2001. On Spectral Clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 14, 849-856.

Poulakis, Z., Valougeorgis, D., and Papadimitriou, C., 2003. Leakage detection in water pipe networks using a Bayesian probabilistic framework. Probabilistic Engineering Mechanics 18, 315-327.

Puust, R., Kapelan, Z., Savic, D. A., and Koppel, T., 2010. A review of methods for leakage management in pipe networks. Urban Water Journal 7(1), 25-45.

Romano, M., Kapelan, Z., and Savić, D., 2011. Real-Time Leak Detection in Water Distribution Systems. Water Distribution Systems Analysis, 1074-1082.

Schaeffer, S.E., 2007. Graph Clustering (survey). Computer Science Review, 27-64.

Scholkopf, B., Smola, A. J., 2002. Learning with kernels. Support Vector Machines, regularization, optimization and beyond. Massachussetts Institute of Technology, USA.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888-905.

Sivapragasam, C., Maheswaran, R., and Venkatesh, V., 2007. ANN-based model for aiding leak detection in water distribution networks. Asian Journal of Water, Environment and Pollution 5(3), 111-114.

Vapnik, V., 1998. Statistical Learning Theory. New York, Wiley.

Xia, L., and Guo-jin, L., 2010. Leak detection of municipal water supply network based on the cluster-analysis and fuzzy pattern recognition. 2010 International Conference on E-Product E-Service and E-Entertainment (ICEEE) 1(5), 7-9.

Xia, L., Xiao-dong, W., Xin-hua, Z., Guo-jin, L., 2006. Bayesian theorem based on-line leakage detection and localization of municipal water supply network. Water and Wastewater Engineering 12.

Zhang, X., Liu, J., Du, Y., Lv, T., 2011. A novel clustering method on time series data, Expert Systems with Applications, 38, 11981-11900.