Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 108 (2014) 235 – 248

Procedia
Social and Behavioral Sciences

AIRO Winter 2013

Improving Analytics in Urban Water Management: A Spectral Clustering-based Approach for Leakage Localization

A. Candelieri [a,b]*, D. Conti [b,c], F. Archetti [a,b]

[a]*Department of Information, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126, Milan, Italy*

[b]*Consorzio Milano Ricerche, Via Cozzi 53, 20126, Milan, Italy*

[c]*Department of Operations Research, University of the Andes, Nucleo La Hechicera, Merida, 5101, Venezuela*

Abstract

Worldwide growing water demand has been forcing utilities to successfully manage their costs. Contemporarily, within an era of tight budgets in most economic and social sectors, it affects also Water Distribution Networks (WDN). So, an efficient urban water management is needed to get a balance between consumer satisfaction and infrastructural assets inherent to WDN. Particular case is referred to pipe networks which suffer for frequent leaks, failures and service disruptions. The ensuing costs due to inspection, repair and replacement, are a significant part of operational expenses and give rise to difficult decision making. Recently, the goal regarding the improvement of the traditional leakage management process through the development of analytical leakage localization tools has been brought to the forefront leading to the proposal of several approaches. The basis of all methods relies on the fact that leaks can be detected correlating changes in flow to the output of a simulation model whose parameters are related to both location and severity of the leak.

This paper, starting from a previous work of the authors, shows how the critical phases of leak localization can be accomplished through a combination of hydraulic simulation and clustering. The research deals with the benefits provided by Spectral Clustering which is usually adopted for network analysis tasks (e.g., community or sub-network discovery). A transformation from a data points dataset, consisting of leakage scenarios simulated through a hydraulic simulation model, to a similarity graph is presented. Spectral Clustering is then applied on the similarity graph and results are compared with those provided by traditional clustering techniques on the original data points dataset. The proposed spectral approach proved to be more effective with respect to traditional clustering, having a better performance to analytically localize leaks in a water distribution network and, consequently, reducing costs for intervention, inspection and rehabilitation.

Keywords: spectral clustering; partition clustering; leakage localization; urban water systems; water distribution network management.

* Corresponding author. Tel.:+39 02- 6448-2184.
E-mail address: antonio.candelieri@unimib.it

doi:10.1016/j.sbspro.2013.12.834

1. Introduction

Water is crucial in human society for both social and economic development (Adams, 2006; Hoekstra, 2006; Jung et al., 2011). Nowadays, water distribution networks (WDN) need to face with a rising demand, dynamic modifications due to change in consumer profiles and often aged assets. Urban WDN management has to deal with these issues by developing and implementing strategies focused on reaching a balance between a maximization of consumer satisfaction and a minimization of efforts in terms of water, energy - and consequently costs - savings.

In particular, water demand is usually affected by time granularities, seasonal influences and climatic conditions (Shvartser et al., 1993; Zhou S.L., et al., 2002; Herrera M., et al., 2010). This dynamic behaviour produces fast variations in pressure and flow within the network, consequently affecting structural components as pipes and junctions of the distributions systems. The most common consequences are breakages and leakages, generating both financial losses and social impacts, i.e., increasing costs related to the increased request for energy to satisfy demand (Puust et al., 2010), replacement and rehabilitation costs and finally, social/health costs related to water quality and diffused infections due to leakages (Puust et al., 2010).

Under these premises, a better water distribution network management is needed to standardize higher levels of efficiency. IWA performance indicators (Alegre et al., 2006) detail the relevance to improve the leakage management process, generally defined on three different steps: *assessment, detection and physical localization* (Preis et al., 2010).

This paper is an extension of previous works developed by the authors and oriented to develop decision support services for leakage localization through a combination of clustering techniques and hydraulic simulation of “leakage scenarios”. The main goal was to identify, according to pressure and flow measurements acquired on the real WDN, the most similar simulated “leakage scenarios” and select a restricted set of pipelines as probably affected by a leak. In this way, physical check can be targeted more effectively and efficiently, reducing time and cost both for inspection and consequent rehabilitation (Candelieri & Messina, 2012; Candelieri et al., in press).

This study deals with the same paradigm but it investigates the benefits provided by Spectral Clustering techniques compared to other classical partitioning strategies (K-means, K-Medoids, etc.). Recently, Spectral Clustering (Luxburg, 2007) has become very popular due to its easy implementation and good results in comparing with classical clustering algorithms, in particular in the fields of graph clustering (Schaeffer, 2007) and network analysis – even social network analysis, such as to identify communities (Aggarwal, 2011). Spectral Clustering-based approaches have been proposed to optimize WDN management taking into account the natural relationship between a WDN and a graph. In (Herrera, 2011) graph/spectral clustering is used to improve leak detection and to screen system vulnerabilities in a WDN of downtown Celaya city (Guanajuato, Mexico), in the same line, (Herrera et al., 2012) address clustering approaches to manage node-pipe in hydraulic sectorization by combining these techniques with multi-agent systems and finally, in (Gutierrez-Perez et al., 2012) spectral clustering supported by a rank function - very similar to PageRank widely used by Google to classify web pages – is performed to get supply groups as a management strategy to vulnerability assessment processes within a WDN.

Differently, our study proposes the adoption of Spectral Clustering in order to compute a possible similarity relationship between different “leakages scenarios” generated and then use this relationship to create a network of leakage scenarios on which Spectral Clustering has been performed.

The paper is structured as follows: section 2 describes material and methods, in section 3 experiments are presented and results are provided. Section 4 closes the paper with conclusions and future works.

2. Materials and methods

2.1 The case study and the available data

The data used in this study are related to a real WDN (H2OLeak Italian project) serving about 2600 users for an average consumption of 0.03302 l/s (each user). The WDN covers a geographical zone of almost 5 km^2; it has more than 45 km of pipes. The network is composed of 931 pipelines and 898 junctions (only one reservoir). A Supervisory Control And Data Acquisition (SCADA) system is used to acquire pressure and flow values at specific monitoring points (1 for flow monitoring at the pumping station and 6 for pressure).

Information about the water distribution infrastructure and data about customer consumption have been used to configure a hydraulic model and then simulate different "leakage scenarios" by placing, in turn, a virtual leak on each pipe, with different severities.

Hydraulic simulation was performed by using the open-source software, EPANET 2.0 from the Environmental Protection Agency (EPA). After each simulation, resulting pressure and flow values, in correspondence of the real monitoring points, have been stored in order to built our leakage scenarios dataset. Clustering techniques have been then applied on this dataset; in particular, the open-source statistical software R (R Core Team, 2013) has been used to perform the clustering algorithms both classical partitioning algorithms (PAM-Clara) and the Spectral ones.

2.2 Building and labeling of leakage scenarios

EPANET 2.0 permits to place leaks within a WDN model. The water flow rate depends on the pressure at the leak location. It follows this mathematical relation:

$$q = Cp^{\gamma} \tag{1}$$

Here, q is the flow rate, p is the pressure, C is the discharge coefficient and γ is the pressure exponent. While wider is the set of leakage scenarios (generated and simulated) more accurate will be the association between the set of pipes affected by a leak and the pressure and flow at the monitoring points. Hydraulic simulation was carried out by following the following strategy: a leak was applied, in turn, on each pipe with C varying from 0.001 to 0.03 with a step of 0.001. So, the whole number of leakage scenarios is given by $n_p * n_c$, with n_p as number of pipelines and n_c as number of possible values for C. The resulting dataset is composed by 27930 instances (931 pipelines * 30 discharge coefficient values, C). This dataset has 9 attributes; the first 7 are computed as the difference between values of each leakage scenario and faultless network: 6 for pressures at the monitoring points and the last one for flow at the pumping system. Attributes 8 and 9 describes ID of pipelines and discharge coefficient associated to each ID; these last two features have been ignored while clustering.

2.3 Spectral Clustering

Spectral Clustering (Luxburg, 2007) has become one of the most popular modern clustering algorithms. It is simple to implement, can be solved efficiently by standard linear algebra software (but it has high computational costs and time consuming in large datasets). Although it has been proposed to solve graph clustering problems in the network analysis domain, it very often outperforms traditional clustering algorithms, such as the k-means algorithm or other partitioning algorithm, when applied on traditional data points datasets. Although the advantages it provides, such as easy implementation and good performance, some authors report disadvantages mainly related to the tuning of parameters (i.e., definition of the similarity/affinity matrix) and the computational effort on large datasets.

In particular, given a set of data points $x_1, ..., x_n$ and some similarity measure $s_{ij} \geq 0$ between each x_i and x_j, traditional clustering approaches identify a partition of the data points into several groups in order to minimize intra cluster similarity and maximize inter cluster similarity. The matrix S is usually defined as affinity matrix and denoted by A.

A possible way of representing the data points consists of building a similarity graph $G = (V, E)$, where vertices v_i are the original data points x_i and edges e_{ij} are weighted by the corresponding s_{ij} of the Affinity matrix (v_i and v_j are not connected by any edge if $s_{ij}=0$). At this point, the problem can be reformulated as a graph clustering task with the aim to identify a partition of the undirected similarity graph such that the edges between different groups have a very low weight (i.e. points in different clusters are dissimilar from each other) and the edges within a group have high weight (i.e. points within the same cluster are similar to each other).

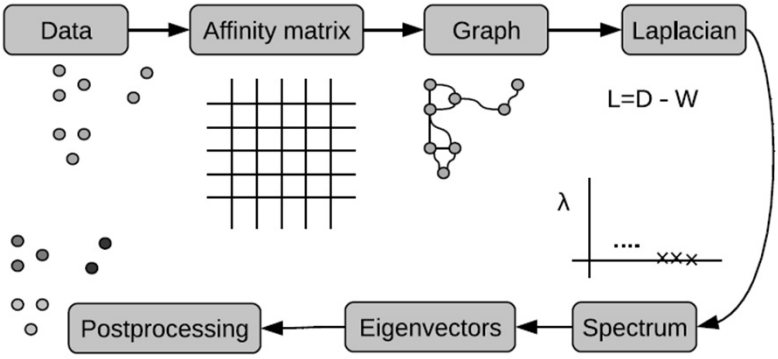

Fig. 1. Application of spectral Clustering to a data points dataset

In our case, the hydraulic simulation of several leakage scenarios were represented as an undirected graph, where each vertex represents a leakage scenario and each edge is weighted by the similarity between pairs of leakages scenarios.

Given an undirected graph with its Affinity matrix A it is possible to compute its Degree matrix D as follows:

$$D(i,i) = \sum_j A(i,j) \tag{2}$$

$$D(i,j) = 0 \, , \; i \neq j \tag{3}$$

The core of Spectral Clustering technique is the graph Laplacian matrix. Different alternative definition exist for this matrix, and graph theory is devoted to their study (Chung, 1997). The (not normalized) Laplacian matrix L is defined as :

$$L = D - A \tag{4}$$

The most important properties of the L matrix are:

- it is symmetric and positive semi-definite;
- its smallest eigenvalue is 0;

- it has n non-negative, real-valued eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.

Many applications use a normalized Laplacian instead of basic Laplacian shown in (4). We have adopted the following normalized Laplacian matrix notation:

$$L^{norm} = I - D^{-1/2} A D^{-1/2} \tag{5}$$

Given a similarity graph with affinity matrix A, the simplest and most direct way to construct a partition is to solve the *mincut* problem. In particular, given two sets of vertices, C_1 and C_2, the objective is to minimize:

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

In the case of bi-partitioning the *mincut* problem is relatively easy to be solved efficiently. In detail, a N-dimensional vector p is used to represent the association of each instance (vertex in the similarity graph) to one of the two possible clusters C_1 and C_2 , where:

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

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

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

Solving this combinatorial problem can be complex for real world networks (from thousand to million or billion of vertices). 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 partitions the graph through the minimum cut (*algebraic connectivity*).

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 (usually known as *K-way graph cuts*). However this approach requires the computation of matrices and eigenvalues, as well as the use of the Fielder vector, for each sub-networks until the desired number of groups is reached.

Another possible schema to solve the K-way graph cuts uses the eigenvectors differently by providing 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 according to the *eigengap* that is the difference between two successive eigenvalues, where eigenvalues are sorted in ascending order.

For example, the K-way 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 (vertexes of the similarity graph, that are initial instances) and l columns (eigenvectors corresponding to the l smallest eigenvalues).

According to the latest approach, the Spectral Clustering procedure adopted in this study consists of the following steps:

- Step 1. “R”, original dataset. This is the $N \times m$-dimensional data that user is interested to cluster. The m features are related to pressure and flow measurements.
- Step 2. “A”, Affinity (or Adjacency) matrix, the $N \times N$ similarity matrix between each pair of instances of R. Different type of similarity measure can be used, such as correlation, Euclidean distance based similarity, kernel functions.
- Step 3. “D”, Diagonal degree matrix, also $N \times N$, formed by summing, for each vertex, the weight of the edges incident on it.
- Step 4. “L”, Normalized and symmetric Laplacian matrix, computed according to (4) and (5).
- Step 5. “U”: A $N \times N$ matrix having the eigenvectors of L as columns.
- Step 6. “U^l”: Select l eigenvectors corresponding to the l lowest eigenvalues (non-zero) in order to define a l-dimensional subspace. U^l is a $N \times l$ matrix obtained from U^l through column selection.
- Step 7. K-means (or other partition clustering) is performed on the rows of U^l (leakage scenarios in the l smallest eigenvectors space).
- Step 8. Evaluation of resulting clusters.

2.4 Experiments design

In this work, experiments have been designed by following the step by step procedure described at the end of section 2.3. However, some additional issues were added in order to create a comparative analysis.

First, a sampling approach was decided in order to avoid high computation and time consuming. From the initial 27930 instances, two samples of size 3000 instances (each one) were extracted randomly from the whole dataset without overlapping. These samples were labeled as “sampling 1” and “sampling 2”.

Second, quality measures regarding the choice in the number of clusters “k” (for clustering purposes) are presented. Silhouette (Rousseeuw, 1987) and Calinski-Harabasz index (Calinski & Harabasz, 1974) were calculated for several values of k in each clustering algorithm over the spectral data. A short definition and interpretation of these two quality measures is given as follows:

Silhouette (Rousseeuw, 1987): For each observation i, the silhouette width $s(i)$ is defined. Put $a(i)$ = average dissimilarity between i and all other points of the cluster to which i belongs (if i is the only observation in its cluster, $s(i) = 0$ without further calculations). For all other clusters C, put $d(i,C)$ = average dissimilarity of i to all observations of C. The smallest of these $d(i,C)$ is $b(i) = min_C \{d(i,C)\}$, and can be seen as the dissimilarity between i and its “neighbor” cluster, i.e., the nearest one to which it does not belong,

$$s(i) = (b(i) - a(i)) / max(a(i), b(i)) \quad (6)$$

Observations with a large $s(i)$ (almost 1) are very well clustered, a small $s(i)$ (around 0) means that the observation lies between two clusters, and observations with a negative $s(i)$ are probably placed in the wrong cluster.

Calinski-Harabasz (Calinski & Harabasz, 1974): Calinski-Harabasz statistic, which is

$$CH(cn)=(n-cn)*sum(diag(B))/((cn-1)*sum(diag(W))) \quad (7)$$

Where, n as number of instances, cn as number of clusters, B being the between-cluster means, and W being the within-clusters covariance matrix. The value of cn which maximizes $CH(cn)$, is regarded as specifying the number of clusters.

Within these previous definitions and remarks, spectral clustering should be performed in our data set as follows:

- Step 1. Obtain samples of the whole dataset (only in case of computing restrictions). Simple-random without replacement is recommended. These samples will be representations of whole dataset. They are the $N \times m$-dimensional datasets in the original space of pressure and flow measurements. In our study N=3000 and m=7.
- Step 2. Create "A" matrix. Affinity matrix is obtained by calculating correlation on each pair of points x_i, x_j in samplings. Give names S_1, S_2, .., S_t (t, number of samples). Here, two matrixes (t=2) of type *"S"* were obtained with a dimension of 3000*3000. Then, assign $A = S - I$ (matrix notation) for each sample.
- Step 3. Compute "D" matrix. Diagonal degree matrix, also of the type $N \times N$ (3000*3000). It is obtained by summing the degree of each vertex and placing it on the diagonal like an almost full connected graph. Here again, two matrixes D will be present in our experiments.
- Step 4. Obtain "L", Normalized Laplacian matrix for each A. Then, eigenvalues and eigenvectors over each L are calculated.
- Step 5. Get "U", matrix of eigenvectors of L.
- Step 6: Choose the number of eigenvalues and criteria to build-up U^l (Gutierrez-Perez et al., 2012). In this study, the four lowest eigenvalues and theirs correspondent eigenvectors were selected to define the new 4-dimensional subspace (U^l). This choice of eigenvalues was taken due to that from all over 3000 eigenvalues, the last four have the biggest difference amongst them. Previous 2995 eigenvalues have minimal difference, almost imperceptible. Eigenvalue number 3000 is zero and it is not taken within spectral analysis.
- Step 7: Clustering
    - Step 7.1. Hierarchical clustering (distance "Euclidean", method "ward") and partition algorithms (k-means and Partitioning around Medoids (PAM)) could be applied for different values of k (number of clusters) over U^l. Silhouette and Calinski-Harabasz index will be used to determine both the best value of k and the algorithm which performs better these quality measures.
    - Step 7.2. Elected algorithm could be applied to "non-transformed dataset" ($N \times m$ dimension) in order to establish benchmarking analysis.
- Step 8. Interpretation of clusters results will be done by comparing spectral clustering results versus traditional clustering approaches (in particular by taking into account the fitness function defined in the next section 3.1).

3. Experiments

Experiments were developed by following the methodological outline explained in section 2.4. Steps 1 to 6 correspond essentially to matrix computing. With U^l matrices step 7.1 was performed to determine the best clustering algorithm which fits the best values for the quality measures with the corresponding "optimal" k (number of clusters) value. This step is supported by simulating clustering techniques for several values of k within different algorithms (hierarchical, k-means and PAM-Clara).

Unsupervised approach was taken at the beginning with hierarchical clustering (distance criterion = "Euclidean", method = "Ward") to determine an initial range for k. Analysis made on dendogram's height suggested k values equal to 4, 5 and 8. Under these premises, simulated clustering with k values from k = 2 to 8 were performed (k-means and PAM) and quality measures were calculated in order to select the best algorithm. Fig. 2 shows an example of this type of simulation by using k-means and Calinski-Harabasz index.

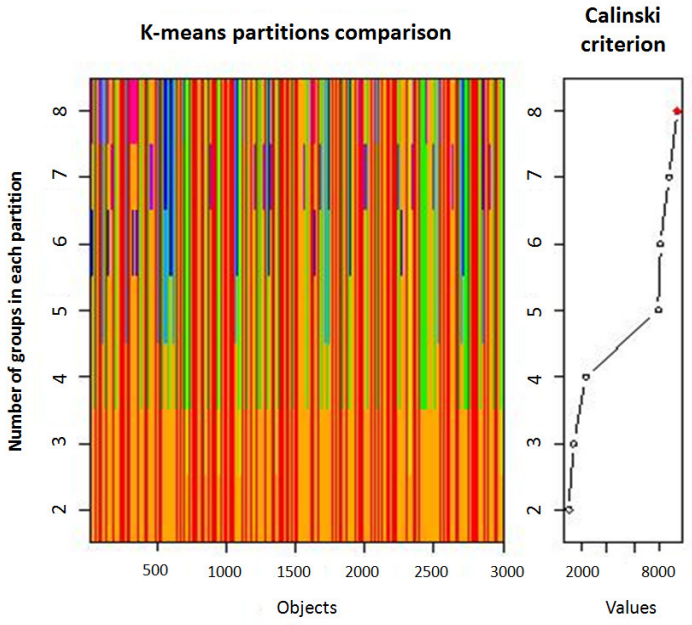

Fig. 2. Detecting optimal k value in K-means algorithm by using Calinski-Harabasz Index

From Fig. 2, it could be seen that k= 8 maximizes the value of the Calinski-Harabasz index. However, values from k= 5 to 7 have similar Calinski-Harabasz indexes (a choice among this range could be also feasible). In order to refine procedure and made a decision, Silhouette criterion is also performed. Fig.3 shows that Silhouette values for k=5 and k=8 are similar and around 0.65 (a good average value).

Values for k = 5 and/or 8 seem to be optimal and are closer to values detected from the first unsupervised approach (hierarchical clustering). So, a previous decision was made, i.e., clustering analysis would be performed by using k=5 and k=8 over the two samples and the non-transformed dataset.

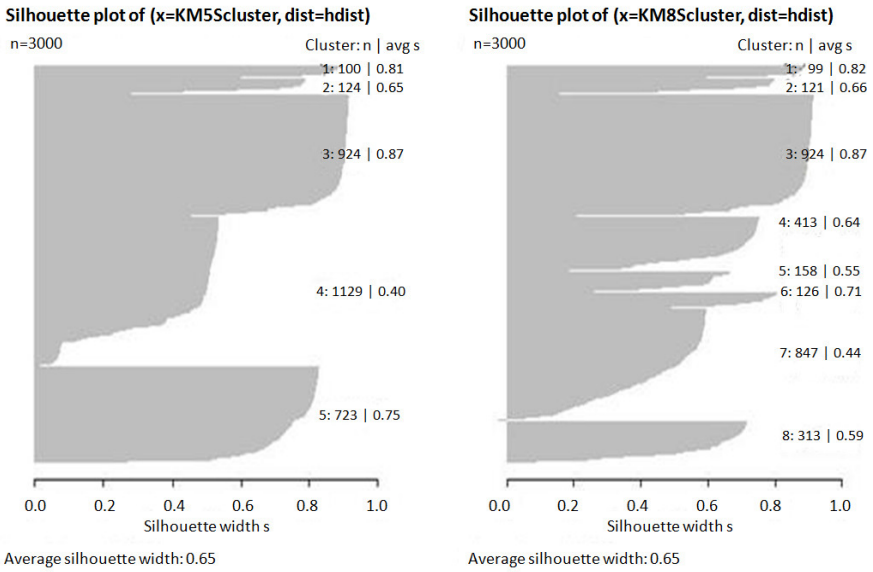

Fig. 3. Detecting optimal k value in K-means algorithm by using Silhouette

Then, a comparative analysis was made by selecting the algorithm which maximized both Calinski-Harabasz index and Silhouette with k=5 and k=8 from three options: hierarchical (cutting dendogram in 5 and 8), K-means and Partitioning around Medoids (PAM). The best performance regarding the two quality measures was found with Partitioning around Medoids. So, PAM (with k=5 and k=8) was used to interpret clustering over spectral data. Also, these algorithms were applied over the non-transformed data to establish comparative analysis.

3.1 Clustering Fitness

While clustering is focused on grouping leakage scenarios by maximizing intra cluster similarity and minimizing the inter cluster similarity (only according to pressure and flow values) the results provided by the different strategies have been evaluated according to the capability to generate a greater number of *localizing* clusters. More in detail, a localizing cluster contains scenarios referred to a limited set of (simulated) leaky pipelines with no respect to discharge coefficient. According to this definition, features 8, ignored in clustering, has been then used for results interpretation.

The procedure was as follows, in each dataset we proceeded to count the numbers of unique pipes from de total number of 931 which contains our case of study (notice that for each dataset, there are pipes which appears repeatedly because each pipe ID has associated with itself different values of discharge coefficients). For sampling 1 and its non-transformed dataset, the total of unique pipes was 903. On the other hand, for sampling 2 and its non-transformed data, this total was 901. Both sampling 1 and 2 have a high numbers of pipes in relation to the whole set of pipelines (931) of the system. In both samplings this representation reaches almost 97% (903/931 and 901/931, respectively). This procedure was repeated to extract unique pipelines from each group obtained from clustering. In particular we have computed the following index:

*Network Coverage Function (%) = (TNPinK / TNPinD)*100* (8)

where *TNPinK* is the total number of unique pipes related to scenarios into cluster K and *TNPinD* is the total number of unique pipes into the whole dataset (sampling1 and sampling2, non-transformed datasets).

The following Fig. 4 shows, on a toy WDN, an example of "good" and "bad" localizing clusters.

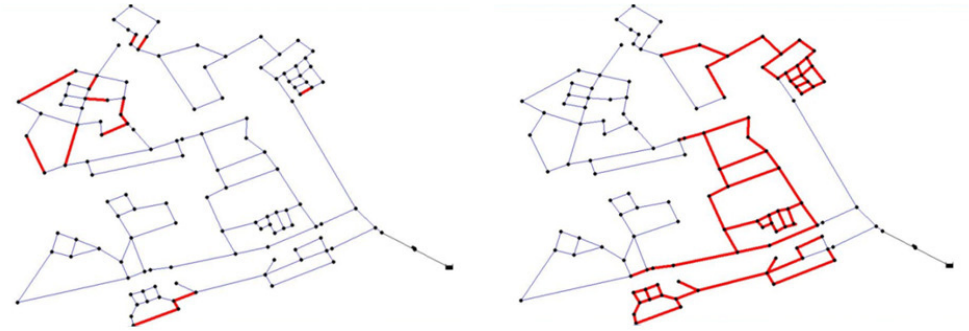

Fig. 4. An example of a "good" (left) and "bad" (right) localizing cluster (highlighted pipelines are those related to scenarios belonging to the cluster)

The next figures and tables summarize our results by showing localizing clusters for each sampling, non-transformed dataset and the clustering algorithm performed, i.e., PAM with K= 5 and 8.

Table 1. Number of pipes (left) and Network Coverage Function (right) for each cluster - Spectral *vs* Non-Transformed Data - Sampling 1

K=5		K=8		K=5		K=8	
Spectral	NTD	Spectral	NTD	Spectral	NTD	Spectral	NTD
315	474	271	426	34,88%	52,49%	30,01%	47,18%
304	473	242	353	33,67%	52,38%	26,80%	39,09%
221	468	217	344	24,47%	51,83%	24,03%	38,10%
36	463	70	336	3,99%	51,27%	7,75%	37,21%
30	448	39	307	3,32%	49,61%	4,32%	34,00%
		33	281			3,65%	31,12%
		31	279			3,43%	30,90%
		30	273			3,32%	30,23%

Table 2. Number of pipes (left) and Network Coverage Function (right) for each cluster - Spectral *vs* Non-Transformed Data - Sampling 2

K=5		K=8		K=5		K=8	
Spectral	NTD	Spectral	NTD	Spectral	NTD	Spectral	NTD
307	496	269	389	34,07%	55,05%	29,86%	43,17%
300	466	239	366	33,30%	51,72%	26,53%	40,62%
229	462	219	346	25,42%	51,28%	24,31%	38,40%
40	445	73	321	4,44%	49,39%	8,10%	35,63%
29	428	39	308	3,22%	47,50%	4,33%	34,18%
		35	301			3,88%	33,41%
		29	281			3,22%	31,19%
		24	280			2,66%	31,08%

Results from Table 1 and 2 are appreciated graphically in Fig. 5 and 6.

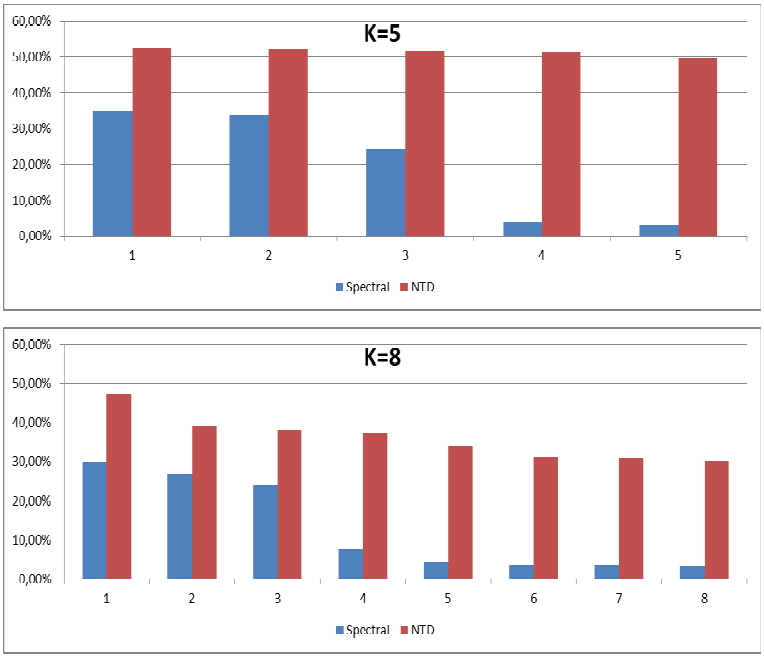

Fig.5. Network Coverage Function for each cluster: Spectral versus Non-transformed data, Sampling 1 (For K=5 and K=8)

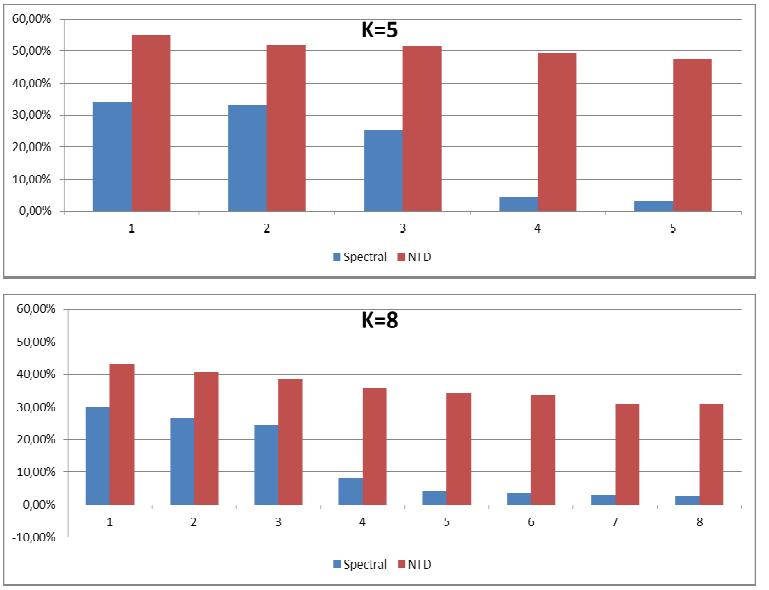

Fig.6. Network Coverage Function for each cluster: Spectral versus Non-transformed data, Sampling 2 (For K=5 and K=8)

In Fig.5 it is possible to realize that clusters obtained through spectral clustering versus traditional are more localizing. Notice that for K=5, in spectral data, there are three groups of pipes with Network Coverage Function lower than 25% and the other two groups have a Network Coverage Function lower than 40%. Meanwhile, for non-transformed data, the Network Coverage Function is higher than 50% in four groups and the last one is over the border of 50% (49.61%), so it is evident that spectral approach is capable to obtain a better leak localization on our simulation scenarios. The same occurs with K=8: Although performances improved for non-transformed data, its performance is absolutely inferior when compared to spectral clustering. Same observation could be extracted from Fig.6 that shows results for sampling 2.

Now, if we consider benchmarking all results for K=5 and K=8 (Fig. 7), it is seen that for both values of K, spectral clustering over the two samplings has a decreasing trend related to localizing index. This decreasing behavior is optimal for our purposes because while more highly localizing clusters more the efficiency in leakage localization (a smaller number of pipes to check implies that costs regarding inspection, rehabilitation or replacement reduce considerably). In addition, note from Fig. 7, that average percentage in spectral clustering is less than 20% which implies that most clustering groups are highly localizing. On the other hand, the performance for classical partition clustering applied directly to non-spectral data is over 50% for K=5 which indicates that cluster groups are not so effective in localizing a leak. For K=8 there is an improvement for the traditional approach, however, its performance continues to be inferior when compared to spectral clustering.

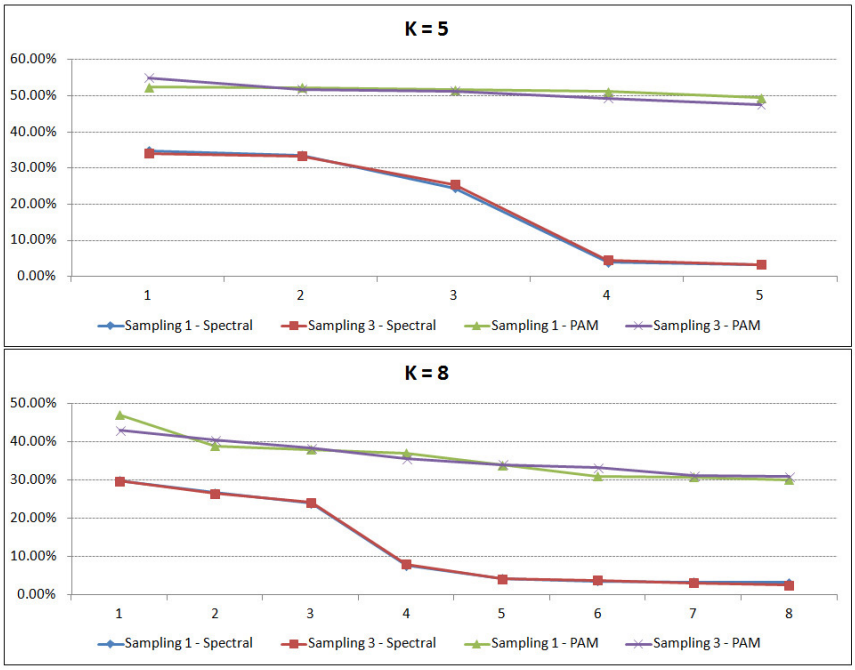

Fig. 7. Network Coverage Function: Spectral Clustering versus PAM, both on sampling 1 and sampling 2, for K=5 and 8, separately

Conclusion & Future works

This paper deals with the adoption of Spectral Clustering approaches to develop a leak localization task based on a combination of hydraulic simulation and clustering. In particular, this work represents an extension of a previous approach using traditional data points clustering and proved to be useful in order to improve the traditional leakage management process in WDN (Candelieri & Messina, 2012; Candelieri et al., in press). The advances proposed by the authors in this paper are related to the transformation of a data points dataset, related to leakage scenarios, into a similarity graph on which is possible to apply Spectral Clustering techniques. Although it has been usually adopted to solve graph clustering problems in the network analysis domain, it proved to outperform traditional clustering algorithms when applied on traditional data points datasets.

Results obtained in this work have confirmed the benefits of Spectral Clustering, offering higher performances with respect to traditional, even if recent, data points clustering approaches. According to these results, the overall approach previously proposed by the authors may be further improved, enabling a more reliable leakage localization and a consequent reduction for time and costs of intervention, investigation and rehabilitation. Moreover, the spectral approach appears to be more effective with respect to those previously proposed by the authors and based on partition clustering approaches proposed in other application domain (Archetti et al. 2006; Fersini et al., 2010).

As future works, authors are planning to investigate strategies in order to identify the best number of cluster k to be used according to the structural properties of the network as well as of the leakage scenarios simulated. Moreover, a more accurate fitness function for clustering will be defined. Finally, high performance computing

solutions, hardware and software, will be investigated in order to use the spectral approach on the overall leakage scenarios dataset, avoiding sampling.

References

Adams, W.M., (2006), The Future of Sustainability: Re-thinking Environment and Development in the Twenty-first Century, *Report of the IUCN Renowned Thinkers Meeting*, 29-31 January 2006.

Aggarwal, C.C., (2011), *Social Network Data Analytics*, Springer New York Dordrecht Heidelberg London.

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.

Archetti, F., Campanelli, P., Fersini, E., Messina, E. (2006), A hierarchical document clustering environment based on the induced bisecting k-means, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4027 LNAI, 257-269.

Calinski, T., and Harabasz, J., (1974) A Dendrite Method for Cluster Analysis, *Communications in Statistics, 3*, 1-27.

Candelieri, A., Messina, E., (2012), Sectorization and Analytical Leaks Localization in the H2OLEAK Project: Clustering-based Services for Supporting Water Distribution Networks Management, *Environmental Engineering and Management Journal, 11(05)*, 3-10.

Candelieri, A., Archetti, F., Messina, E. (in press), Analytics for Supporting Urban Water Management, In *Proceedings of International Conference ECOIMPULS 2012 - Environmental Research and Technology*.

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

Fersini, E., Messina, E., Archetti, F., (2010) Multimedia summarization in law courts: A clustering-based environment for browsing and consulting judicial folders, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6171 LNAI, 237-247.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Math. J.*, 23, 298–305.

Gutierrez-Perez, J., Herrera, M., Izquierdo, J., Perez-Garcia, R., (2012), An approach based on ranking elements to form supply clusters in water supply networks as a support to vulnerability assessment, *In Proceedings of International Congress on Environmental Modelling and Software*, iEMSs 2012, Leipzig, Germany

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

Herrera M., Torgo, L., Izquierdo, J., Perez-Garcia, R., (2010), Predictive models for forecasting hourly urban water demand, *Journal of Hydrology*, *387*, 141-150

Herrera, A.M., (2011), Improving water network management by efficient division into supply clusters, PhD Thesis, Universitat Politècnica de València, Spain.

Herrera, M., Gutierrez-Perez, J., Izquierdo, J., Perez-Garcia, R., (2012), Combining multiple perspectives on clustering: Node-pipe case in hydraulic sectorization, *Int. J. Complex Systems in Science, vol. 2(1),* 17-20.

Hoekstra, A.Y., (2006), The global dimension of water governance: Nine reasons for global arrangements in order to cope with local water problems, *Value of Water Research Report Series No. 20*, UNESCO-IHE Institute for Water Education, Delft - The Netherlands.

Jung, N.C., Popescu, I., Price, R.K., Solomatine, D., Kelderman, P., Shin, J.K., (2011), The use of the A.G.P. test for determining the phytoplankton production and distribution in the thermally stratified reservoirs: The case of the Yongdam reservoir in Korea, *Environmental Engineering and Management Journal*, *10*, 1647-1657.

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

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

Preis, A., Allen, M., Whittle, A.J., (2010), On-line hydraulic modeling of a Water Distribution System in Singapore, *Water Distribution System Modeling Issues*, 1336-1348

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

R Core Team (2013), R: A language and environment for statistical computing, R Foundation for statistical computing, Vienna, Austria.

Rousseeuw, P.J., (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

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

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

Shvartser L., Shamir, U., Feldman, M., (1993), Forecasting hourly water demands by pattern recognition approach, *Journal of Water Resources Planning and Management*, *119(6),* 611-627

Zhou S.L., McMahon, T.A., Walton, A., Lewis, J., (2002), Forecasting operational demand for an urban water supply zone, *Journal of Hydrology, 259*, 189-202