Jul 1st, 12:00 AM

Simulation and Machine Learning Strategies for Enabling Integrated Water Resource Management: H2OLEak Project

Antonio Candelieri
Enza Messina
Alberto Malacrida

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference


This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Simulation and Machine Learning Strategies for Enabling Integrated Water Resource Management: H2OLEak Project

Antonio Candelieri¹, Enza Messina¹, Alberto Malacrida²
¹Department of Information, Systems and Communication, University of Milano-Bicocca, Milan, Italy
antonio.candelieri@unimib.it, enza.messina@unimib.it
²GESP, Geographic Information Systems, Milan Branch, Milan, Italy
malacrida@gesp.it

Abstract: In this paper the main Decision Support functionalities proposed and developed within the H2OLEAK research project are presented. H2OLEAK, co-founded by Regione Lombardia in Italy, is aimed at designing and developing an innovative technological platform for supporting a rational and integrated management of urban water distribution systems. It integrates already available and robust technological solutions, such as Supervisory Control And Data Acquisition (SCADA) systems, Geographical Information Systems (GIS) and Business Intelligence tools, with advanced analytical methodologies to support managers in their decision making activities, enabling prompt and proactive actions that may reduce costs while guaranteeing high customers satisfaction. In particular, computational approaches proposed to address three main problems are described: (i) automatic districts identification to obtain the “optimal” partition of a water distribution system into virtually independent sub-networks, (ii) computational localization of leaky pipelines through the analysis of flows and pressures measured at the entry points of each district and (iii) regression models for estimating the loss intensity of the leak to improve localization effectiveness by further reducing the set of pipelines to be checked physically.

Keywords: Water Resource Management, Sectorization, Leak Detection, Simulation, Clustering

1 INTRODUCTION

The main purpose of a water distribution system is to satisfy customers demand guaranteeing a reliable service. variations in customers demand occur at any time windows (e.g. hourly, daily, weekly) and increase the volatility of flows and pressures, with a consequent mechanic strain for pipelines. Fast demand variations, hydraulic strain, structural modifications, characteristics of the components adopted in expansion and rehabilitation − needed when significant changes in demand distribution occur −, type of ground, climate and weather, all contribute to deteriorate water distribution systems, generating breakages and leakages. As reported by Puust et al. [2010] the main drawbacks generated by leakages are: (i) water and financial losses, due to “extra” energy needed to guarantee a high customer satisfaction despite leaks; (ii) infrastructure and third
party damages; (iii) impacts on health due to low water quality and spreading infections.

During the World Water Day, on 22th March 2011, the Italian National Institute for Statistics, ISTAT [2011] reported that, in average, about 40% of water entered into the Italian water distribution systems is lost, with peaks higher than 80% in the South and the islands (data refer to water consumption in 2008).

The H2OLEAK project (www.h2oleak.it), co-founded by Regione Lombardia in Italy, was devoted to design and develop an innovative technological system for assisting administrators in an Integrated Water Resource Management (IWRM). This is achieved through the integration of already available robust technological solutions, which effectively enables the network monitoring and control:

- a SCADA system to monitor pressures and flows as can be found in Dobriceanu et al. [2008] and Robles et al. [2009];
- a Geographical Information System (GIS) to store data with correspondent location, as shown by Wang et al. [2008] and Yan et al. [2008];
- a Business Intelligence suite, namely QlikView (http://www.qlikview.com/), to assist managers in anomaly detection and leakage assessment;
- a freeware for hydraulic simulation and analysis of pressurized networks, namely EPANET 2.0 (http://www.epa.gov/nrmrl/wswrd/dw/epanet.html).

Moreover, data-driven machine learning based services have been developed for assisting administrators to:

- Identify suitable partitions of the system into virtually independent districts, by using the set of installed valves and according to some preferences that can be set up by users;
- Identify the most probably leaky pipelines enabling prompt and targeted interventions.

The rest of the paper is structured as follows: in the sections 2, 3 and 4 the proposed approaches for automatic district identifications, computational localization of leaks and regression models for discharge coefficient estimation are reported, respectively. Results on a real case study are reported in section 5. Limitations and future works conclude the paper.

2 AUTOMATIC SECTORIZATION SERVICE

The division of a water distribution system in a set of virtually independent sectors, usually defined District Metering Areas (DMA), may significantly make more easy and effective the management of the entire system - in particular when it is wide and intricate - by enabling monitoring and control on narrow sub-networks. As reported in the following, sectorization may also improve leakage management, allowing managers to detect anomalies within specific portions of the system.

In H2OLEAK the districts identification has been defined as the problem of identify an optimal partition of the network nodes into $k$ groups, where any district can be isolated from the others, temporary or permanently, through the installed set of valves. According to specific needs, the administrators can set:

- Number $k$ of desired districts;
- Relative relevance between obtaining homogeneous districts (in terms of demand, level and geographical distance of nodes) and minimizing the number of valves needed for sectorization, namely connectivity;
- Relative relevance of three different factors involved in the computation of intra-district homogeneity: demand, level and location of nodes.

Although the goal of sectorization is to identify a partition of nodes, the problem has a specific constraint related to the “connected” structure of the network (i.e., a non-oriented graph) and has to be faced as a graph clustering task. As reviewed in Schaeffer [2007], many approaches have been proposed for this kind of problems,
depending on specific goals and, essentially, for edges- and nodes- related aggregation, separately. Herrera et al. [2010] proposed a semi-supervised learning approach to identify districts. We propose to solve the sectorization problem by constrained agglomerative clustering, where a connectivity constraint restricts the merge step only to the clusters that are linked by at least a pipeline with valve. A specific requirement of the case study, described in section 5, was that the set of valves is fixed a priori and cannot be changed. This allowed us to cluster, in the same district, all the nodes linked by a valve-less pipeline. However, the number of "structural" districts obtained may be still too high. Thus, the strategy starts from this "structural" partition and, at each iteration, identifies and merges the pair of districts with the objective of minimizing the connectivity while maximizing the intra-cluster homogeneity until the desired $k$ districts are obtained. These are clearly two conflicting goals therefore we need to define their relative relevance in order to reach a suitable trade-off.

3 LOCALIZATION OF LEAKS VIA SIMULATION AND CLUSTERING

Leakage management is a multi-step process, as reviewed by Puust et al. (2010). The first activity is the assessment, aimed at estimating the entity of losses without referring to their possible location. Common approaches evaluate the overall water balance in the network [AHL 2006] and usually consist of a combination of equipment (such as, Leak Reflection Method, Inverse Transient Analysis, Impulse Response Analysis, Transient Dumping Method and Frequency Response Method) and non-equipment based techniques. Among the non-equipment based methods, the 24 Hours Zone Measurements (HZN) technique is performed by temporary isolating portions of the network, while Minimum Night Flow (MNF) analysis is based on the acquisition of flow and pressure measurements during the early morning hours, usually 02:00 to 04:00, as shown in Liemberger et al. [2004]. Over night customers demand is usually near to zero and rising trends over days can be associated to water losses. These two techniques proved to be more effective when applied to single DMAs, as reported in Behzadian et al. [2009], providing leakage detection capabilities, that is the recognition of leaks within a specific portion of the network, as also shown in Covas et al. [2006] and Garcia et al. [2006].

Detection is the second phase of leakage management, and several approaches have been proposed, based both on statistics and machine learning methods, as reported by Buchberger et al. [2004] and Mashford et al. [2009].

Physical localization of the leaks is the last activity of leakage management. Widely adopted methods are based on:

- analysis of acoustic signals, such as vibration sensors and hydrophones, as shown by Muggleton et al. [2004],
- application of ground penetrating radar (GPR), as reported by Farley [2008],
- leak noise correlators, as reported by Muggleton et al. [2005] and
- gas injection, as shown by Farley et al. [2003].

These methods are usually reliable in localizing a leak on a specific pipeline, but are strongly dependent on the experience of the operator and are impracticable whether a reliable detection is not preliminary performed.

The QlikView Business Intelligence tool has been integrated in H2OLEAK to allow managers to visualize and analyze pressures and flows data over time, and support them in the assessment and detection of a leakage. Once a leakage in the network is detected, the costly task of the leak localization must be faced, in order to repair the pipeline and restore the network. A specific decision support service has been implemented for performing this challenging task through a reliable and effective preliminary computational localization of leaky pipelines which reduces costs and time of the physical leak localization. This service involves two processes:

- generation of several leakage scenarios, simulated through EPANET;
clustering of scenarios according to the “effect” produced by each (simulated) loss (i.e., pressure and flow variations at the monitoring nodes).

The "emitter" element of EPANET has been used to emulate leaks: the associated flow rate of the loss depends on the pressure at the node according to the formula $q = Cp^y$, with $q$ the flow rate in the correspondent pipeline, $p$ the pressure at the emitter, $C$ the discharge coefficient (litres per second per meter) and $y$ the pressure exponent. Pressure coefficient differs according both to geometry of the orifice and material of the pipelines, as reported by Greivnstein et al. [2005] and Van Zyl et al. [2007]: 0.5 is generally used to simulate circular orifices. Each leakage scenario is obtained by introducing, in turn, a leak (emitter) on a single pipeline of the faultless network model. Leaks of different intensity can be simulated by varying discharge coefficient, usually from 0.001 to 0.03 with a step of 0.001, for each pipeline, as in Mashford et al. [2009]. The instances (leakage scenarios) generated by this procedure is number-of-pipelines * number-of-discharge-coefficient-values.

Features characterizing a leakage scenario are pressure and flow variations at the monitoring points with respect to the faultless network. Although EPANET uses a quasi-steady model, it was suitable to this study because data were related to daily aggregated consumptions within a pressurized network. Other hydraulic simulation software should be considered in “dynamic” settings (e.g., water distribution systems equipped with advanced metering devices).

To associate pressure and flow variations to a limited set of elements, a fitness function based on the number of clusters that have few pipelines (ignoring discharge coefficient) has been defined, producing the following four type of clusters:

- **Highly Localizing** - having less than 25% of pipelines of the network;
- **Average Localizing** - having between 25% and 50% of pipelines;
- **Poorly Localizing** - having between 50% and 75% of pipelines;
- **No Localizing** - having more than 75% of pipelines.

The following clustering approaches have been initially adopted with the aim to identify the most promising one:

- **Farthest-first**, reported by Hochbaum et al. [1985] and Dagstupa et al. [2002], provided by WEKA, (http://www.cs.waikato.ac.nz/ml/weka/);
- **Agglomerative** that, iteratively, merges the two clusters having minimal inter-clusters distance;
- **Induced Bisecting k-means**, described by Archetti et al. [2006] and Fersini et al. [2010], that has been proposed to overcome drawbacks of the **Standard Bisecting**, proposed by Steinbach et al. [2000] and Savaresi et al. [2001].

All the clustering methods used Euclidean distance.

As reported in section 5, these clustering methods provided similar performances, therefore, we have proposed an ad-hoc extension of the Induced Bisecting, namely "In-deep Bisecting". The idea is to prefer the clusters generation order rather than the value of intra-cluster distance when a cluster has to be selected for splitting. In detail, the In-deep Bisecting algorithm works as follows:

**Step 0:** Initialization: all the instances are associated to the same cluster $S$;
**Step 1:** Insert the initial cluster $S$ in the clusters generation queue $Q$;
**Step 2:** Denote with $S_c$ the first cluster in $Q$ and remove it from $Q$;
**Step 3:** Split $S_c$ in $S_1$ and $S_2$ by selecting the two instances in $S_c$ having maximum distance and denote them as centroids of $S_1$ and $S_2$;
**Step 4:** Place remaining instances of $S_c$ in $S_1$ or $S_2$ according to the distance to the correspondent centroid;
**Step 5:** If both intra-cluster distance and number of instances of $S_2$ are higher than two prefixed thresholds than put $S_2$ in $Q$;
**Step 6:** If both intra-cluster distance and number of instances of $S_1$ are higher than two prefixed thresholds, then update $S_c$ with $S_1$ and go to step 3 else go to step 2;

**Step 7:** If the number of current clusters is equal to the desired one or $Q$ is empty STOP else go to step 2.

After a leakage is detected and assessed, pressure and flow variations, computed with respect to the simulation of the faultless network, are compared to centroids obtained through clustering, in order to identify that most similar. Since any instance represents a leak on a pipeline, the pipelines belonging to the identified cluster are candidate as leaky. The number of pipelines to check by equipment based methods can be further restricted by considering the intersection between the identified cluster and the district indicated by traditional approaches such as MNF analysis. With respect to this, it is clear the importance to propose a clustering method that generates manly highly localizing groups.

### 4 IMPROVING LOCALIZATION THROUGH A REGRESSION MODEL FOR DISCHARGE COEFFICIENT ESTIMATION

Although the proposed In-deep Bisecting method proved to be appropriate enough to identify the highest number of highly localizing clusters, a regression model has been learned to estimate the discharge coefficient of the leak and has been used to improve the efficiency of physical localization. In detail, the regression model works in parallel to clustering - on the same features set - with the aim to estimate the intensity of the loss. After the cluster with the centroid most similar to real flow and pressure variations is identified, pipelines belonging to that cluster are ranked/selected by comparing simulated and regressed discharge coefficients. Least Median Square Linear Regression algorithm (in WEKA) proved to be reliable enough to this aim.

### 5 CASE STUDY

H2OLEAK has been tested on a real case study consisting of a water distribution system of a little town (13km²), with level ranging from 107 to 118.9 meters. The pressurized network satisfies the demand of more than 6300 citizens through about 45 km of pipelines. The number of pipelines in the network model are 931. It is important to note that in Italy water consumption is usually accounted for building and not for single user (flat), thus the number of users (2600) is lower than citizens. The case study is equipped with many valves (about 200) that may yield to different possible sectorizations. For the deployment of monitoring devices and the initial validation of the technological platform, a “manually” defined division in four districts was applied.

#### 5.1 Computational Results on the Automatic Sectorization

Since In-deep Bisecting provides a heuristic solution to the districts identification problem, several parameters configurations have been tested. In the following Figure 1 significant results are reported.

The most relevant anomaly has been obtained by setting to 0 the relative relevance of homogeneity (relevance of connectivity is 1 minus that of homogeneity), resulting in the highest number of pipelines with valve linking different districts. This is essentially due to the heuristic nature of the approach; however, the aim of the service is to assist managers in making decisions and not to operate automatic actions. Respect to this, network managers may perform multiple analyses (i.e., that reported in Figure 1) to identify the most suitable sectorization or modify the proposed one according to their experience and knowledge.
Multiple analyses should be performed any time relevant modifications in customers demand or system’s structure occur. With respect to the heuristic approach proposed by Izquierdo et al. [2010], based on multi-agents, our strategy is independent on the number of supply sources.

\[ \text{Figure 1. Number of pipelines with a valve linking different districts, depending on the relevance of homogeneity for a 4 and 10 districts sectorization, separately.} \]

### 5.2 Computational results on leak localization

For experimental validation we have adopted a setting based on a 6 districts partition, with pressure monitored at the entry point of each district. Flow variations are measured only at the pumping system, unique in the case study: no further tanks are installed. Thus, scenarios are represented by seven features. The total number of scenarios that have been generated is given by \( \text{number-of-pipelines} \times \text{number-of-discharge-coefficient-values} = 931 \times 30 = 27930 \) in our case. In Figure 2 and 3 the number of highly, average, poorly and no localizing clusters is reported, with respect to the four clustering methods and different level of aggregation: 50, 100, and 150 clusters, respectively.

\[ \text{Figure 2. Results for 50 and 100 desired clusters} \]
Figure 3. Results for 150 desired clusters

It is possible to note that the In-deep Bisecting provides the highest number of localizing clusters when 150 groups are achieved. This desired behaviour becomes more relevant while aggregation process goes ahead: more highly localizing clusters are identified at the expense of those less localizing, with respect to the other approaches. Authors are aware of a drawback of the proposed approach: at Step 3 (splitting), the two clusters $S_1$ and $S_2$ are determined according to the two most far instances in $S_c$ and the assignment centroid-to-cluster is performed depending on the order of instances into the dataset, affecting, consequently, the order in which clusters are added into $Q$. However, no significant variations in the number of highly, average, poorly and no localizing clusters were obtained by performing random shuffles of the dataset.

5.3 Regression approaches for estimating discharge coefficient

Least Median Squared Linear Regression proved to be effective enough to predict discharge coefficient of the leak depending on pressure and flow variations. Model is the following; the algorithm automatically removes a monitoring node:

$$dc = -0.0123\Delta P_1 + 0.0025\Delta P_2 - 0.0026\Delta P_3 + 0.0023\Delta P_4 + 0.008\Delta P_5 - 0.081\Delta F$$

where $\Delta P_i$ is the pressure variation at the $i$-th monitoring node and $\Delta F$ is the flow variation at the pumping system, with respect to the faultless network. A 10-fold cross validation provided a high correlation coefficient ($0.9997$) and low error: Mean Absolute Error = $10^{-4}$; Root Mean Squared Error = $2.10^{-4}$; Relative-Mean Absolute Error = $0.8764$%; Root Relative Mean Squared Error = $2.5368$%.

6 CONCLUSIONS

In this paper some computational methods have been presented to implement the main decision support functionalities of an innovative technological platform for supporting a rational and integrated management of water distribution systems. In particular, an agglomerative clustering procedure for automatic districts identification, a combined simulation-clustering approach for leak localization and a regression model to further improve localization effectiveness have been proposed. As shown by the results, preliminary validation both on simulated and real data proved that the proposed methods are more than promising. The districts identification service assists managers to define a suitable sectorization of the network making easier its monitoring and control; the computational localization of leaks permit to identify a set of probably leaky pipelines, reducing the time and costs of intervention and rehabilitation.
Nevertheless some further issue should be addressed. In particular, efficient procedure for providing scenarios with multiple simultaneous leaks should be investigated: current model (clusters set) is based on simulation of single-leak scenarios that could not be sufficient enough to correctly localize multiple leaks. However, the number of generated scenarios relevantly affects the efficiency of our computational localization of leaks; in order to maintain acceptable the time for clustering scenarios, techniques of instances reduction should be also explored. However, it is important to note that the proposed approach is general and independent on the scenarios generation procedure.

REFERENCES


ISTAT, Giornata mondiale dell’acqua, le statistiche dell’ISTAT (trad. World Water Day, the statistics of ISTAT), 2011. (Online: http://www3.istat.it/salastampa/comunicati/non_calendario/20110321_00/testointegrale20110321.pdf)


Schaeffer, S.E., Graph Clustering, Computer Science Review I, 27-64, 2007.


