International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

An approach to water supply clusters by semi-supervised learning

M. Herrera [a], **S. Canu**[b], **A. Karatzoglou**[b], **R. Pérez-García**[a] **and J. Izquierdo**[a]

[a]*IMM–Universidad Politécnica de Valencia, Camino de Vera s/n Edificio 5C 46022, Valencia, Spain (mahefe@gmmf.upv.es, rperez@gmmf.upv.es, jizquier@gmmf.upv.es)*

[b]*LITIS–INSA de Rouen, Avenue de l'Université 76801, Saint-Etienne du Rouvray, France (stephane.canu@insa-rouen.fr, alexis@ci.tuwien.ac.at)*

Abstract: The rational distribution of water in a water supply network (WSN) is a complex problem, especially for systems of large scale. Its complexity is continually increasing from the point of view of technical management. The division of WSN into hydraulic zones is a partition of the supply network into subsystems with controlled inputs and outputs, building smaller independent networks. This solution is a strategic option used in many cities worldwide to control and operate their systems seeking to improve the WSN management, working with each part as a whole. Looking for leaks, detecting water distribution anomalies or carrying out rehabilitation plans, are instances of the aspects that can be technically improved by this reduction of the inspection area. For these reasons, it is important to design the hydraulic zones structure in some optimal way. In this paper, we propose a semi-supervised learning to approach it. To do it we add the different supply constraints to the adjacency matrix of the graph and then gathering the reality of the hydraulic zones in a single matrix. The next step splits the network, applying to it a spectral clustering algorithm. This methodology offers an adequate solution to the hydraulic zones paradigm through clusters that allow the conditions for the zones to become small quasi-independent water supply networks.

Keywords: Kernel methods; clustering; graph partitioning; decision support systems; water supply.

1 INTRODUCTION

Current management strategies, used by companies, hinge on the need of accurately capturing derived infrastructure data sets. Real water distribution systems may consist of thousands of consumption nodes interconnected by also thousands of lines and the necessary elements to feed the network. Most of the times these networks are not the outcome of a single process of design. They are the consequence of years of history giving anarchic response to continually rising new demands. As a result, their layouts lack a clear structure from a topological point of view. Consequently, as other complex systems, water supply networks demand deeper hydraulic knowledge to operate and to carry out tasks of maintenance, guaranteeing the quantity and the regularity of the supply to the final costumer. The division of a network into District Metered Areas (DMA[1]) follows a divide and conquer strategy that splits the large highly interconnected distribution network into smaller sub–networks. These smaller networks are virtually independent and are fed by a prefixed number of sources. This independence can be physically enforced in a number of ways. For instance, by closing valves in existing pipes, by sectioning existing pipes or by introducing new pipes that redistribute the flow.

[1] In this paper we will also refer to DMAs as hydraulic zones.

To achieve such a relevant network division, it is necessary to take into account a number of characteristics guaranteeing the correct layout and dimension of the hydraulic zones. Since this division splits the WSN, simulations of the mathematical model with EPANET by Rossman [2000] or any other hydraulic analysis tool, is required to obtain a good and valid design compatible with the original network purpose. In other words, the division of the network involves some reduction of the water pressure and, consequently, a sufficient number of simulations to guarantee a minimum service pressure at all demand nodes of each hydraulic zone must be performed. In addition, personnel qualification is crucial for better exploiting the network under the new scenario of the network being divided into DMAs. Water only enters to supply a DMA through one or a reduced number of sources. As a result, even a small breakdown or a slightly wrong maneuver at a control point can leave without water the whole sector. It is thus necessary to design by-passes at the control points and to look for alternatives to prevent such eventualities.

From a classical perspective, the division of a water supply network into DMAs is used with the target of leak control like Covas and Ramos [1999], since it helps maintaining a permanent pressure control system. This is de main reason for IWA [2007] to recommend a DMA size between 500 to 3000 service connections (Hunaidi and Brothers [2007]). Nevertheless, this target has become more ambitious recently, and AVSA [2009] incorporated new actions such as: carry out audits to know the hydraulic efficiency, characterize the demand curve, detect frauds or diverse errors of measurement, reduce maintenance costs...

Following the proposed line of considering a DMAs as virtually autonomous networks, that is to say, scaled WSNs endowed with their own facilities to be handled, a complete management of each hydraulic zone of the WSN can be better developed. Also, a new approach to the operation and control of the whole network can be obtained. As a result, the target may be changed, allowing weighting global and local needs. This paper aims to be a contribution to this goal by proposing a technique that efficiently builds hydraulic zones, based on a kernel spectral clustering approach.

Traditionally, the development of DMAs has been strongly empiric, based on technical experience and with very few scientific contributions; in this sense, it is necessary to quote the contributions of United Kingdom Water Industry Research (UKWIR [1999]) and of the guidance notes of the International Water Association (IWA [2007]). Recently, new approaches have been presented for DMA based on conceptual and scientific frameworks. Pioneering work by Hunaidi [2005] uses periodic acoustic surveys in a DMA. Tzatchkov et al. [2006] applies graph theory to establish the division into DMAs. Some of the authors of this paper have obtained interesting results about DMA based in Multi–Agent Systems in Izquierdo et al. [2009]. However, it is necessary to work more in depth regarding hydraulic zones. Our aim is to work with all the available information about the WSN to improve the results obtained until now. Then, the graphical structure of the network and the weighting values of the different variables of the WSN will be taken into account. The goal is to open a new research line to obtain accurate, efficient and robust results about the construction of these hydraulic zones.

The main challenge is to define the kernel matrix (Scholkopf and Smola [2002]) that captures the semantics inherent to the graph structure but, at the same time is reasonably efficient for evaluation. In the first instance, the affinity graph matrix is transformed into a kernel matrix, carrying out the correspondence kernel abstraction of the essential characteristics of the WSN. Next, spectral clustering techniques are applied to this new matrix. Finally, graph-based semi-supervised learning methods (Zhu et al. [2006]) are conducted. These learning methods can be viewed as imposing mechanism of smoothness conditions on the target function with respect to a graph representing the data points to be labeled.

The idea of constructing kernels on graphs (i.e., between the nodes of a single graph) was first proposed by Kondor and Lafferty [2002], and extended by Smola and Kondor [2003]. In this paper we propose an application of this idea to the real case of the WSN of Celaya (Guanajuato, Mexico).

This paper is organized as follows. Section 2 proposes a kernel space approach to performance

WSN data. Section 3 introduces the clustering process as a criteria for the division of the WSN into hydraulic zones. In next section we apply this methodology to a real case. Finally, in Section 5 we comment some conclusions and future research lines.

2 KERNEL ABSTRACTION OF WATER SUPPLY NETWORK DATA

The starting point of our proposal to create DMAs into a WSN is to take into account all the available information of the network. This information will be available as input matrices. Next we see the construction and subsequent treatment of them.

First of all, a WSN must be considered as a particular graph including geographical and connectivity information. Then, we start by building the affinity matrix associated to a WSN. The next correspondences must be considered. First, graph nodes are the consumption points of the WSN and their weights are their water demands. Second, graph links are the pipes of the WSN and their weights are the diameters of these pipes. Using this information an affinity matrix of the graph adapted to a WSN needs can be obtaineding.

We may also be interesed in contemplating the use of other information: using different constraints and information of the water supply in form of dissimilarity matrices. In particular, geographic information of the consumption nodes can be added by using the symmetric matrix that contains the distances between the nodes. Other possibility is to perform a dissimilarity matrix using the elevations of these nodes. Of course, it is possible to add as many matrices as information categories are available.

After building the matrices of input information, they are transformed into so–called kernel matrices. This process proposes an integration of the input data, offering a regulated form to dispose the information and an adecuated operation space. To do it, data are scaled between 0 and 1. Then, a diagonal of 1's is plugged on the diagonal of each matrix. Next, matrices are mirrored through their diagonals to turn them symmetric (the input matrices are triangular).

There are two key properties that a kernel function must meet (Scholkopf and Smola [2002]). Firstly, it should capture the measure of similarity approximate to the particular task and domain, and, secondly, its evaluation should require significantly less computation than it would be needed in an explicit evaluation of a corresponding feature mapping.

Furthermore, as the sum of kernel matrices is another kernel matrix, we propose to build an accumulative matrix, which is the weighted sum of the normalized dissimilarities in the different characteristics of the data (1). In our clustering algorithm (see Section 3) we are interested in representing, in addition, the *a priori* information about the graph (previous DMA designs, i.e.).

$$K = \lambda_A K_A + \sum_{i \in I} \lambda_i K_i \tag{1}$$

K is the kernel matrix for clustering, K_A is the kernel matrix related to the affinity graph and $K_i, i \in I$, is the matrix associated to the inputs of our interest in the process of building hydraulic zones. Finally, λ_A and $\lambda_i, i \in I$, are the weights entering the linear combination. Starting from the information arranged in a suitable kernel matrix we can apply skills of analysis as the next spectral clustering.

3 CLUSTERING PROCESS

This section introduces a number of methodologies that are related with graph clustering and the kernel spectral clustering paradigm. We present them orientated to the applications of our interest, specifically to solve a WSN zoning problem.

3.1 Clustering objective function

The network structure of data is the main reason to approach our problem from the point of view of Graph Clustering (Shi and Malik [2000]). The input is assumed to be a graph $G = (\nu, \epsilon, A)$, where ν is the set of vertices, ϵ is the set of edges and A is the edge affinity matrix. A_{ij} represents the edge-weight between vertex i and j (Kulis et al. [2005]).

Let $\mathcal{A}$, $\mathcal{B}$ two sets of links in the graph. Let $(\mathcal{A}, \mathcal{B}) = \sum_{i \in \mathcal{A}, j \in \mathcal{B}} A_{ij}$ be the links. Let us consider degree($\mathcal{A}$)=$links(\mathcal{A}, \nu)$. We seek a k-way disjoint partition to minimize a particular objective (see Table 1). `Min Cut` objective function can be solved efficiently but the partition may leave isolated vertices to become clusters. This problem is solved with `Ratio Cut`, which creates more balanced clusters. `Normalized Cut` maximizes connectedness inside clusters and disconnectedness between clusters at the same time.

Table 1: Examples of graph clustering objectives

Name	Objective function
Min Cut	$\sum links(\nu_i, \nu_i^c)$
Ratio Cut	$\sum \frac{links(\nu_i, \nu_i^c)}{\lvert\nu_i\rvert}$
Normalized Cut	$\sum \frac{links(\nu_i, \nu_i^c)}{degree(\nu_i)}$

In a WSN a `Min Cut` objective function has the sense of looking for a minimum number of isolation valves able to divide the network.

3.2 Spectral Clustering

Spectral Clustering is a relatively new paradigm on clustering and a really interesting alternative to common methods. We will use here the next version of Spectral Clustering algorithm proposed by Ng et al. [2001]:

Given a set of points $S = \{s_1, ..., s_n\}$ in $\mathbb{R}^l$ that we want to cluster into k subsets:

1. Build the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = exp\left(-\frac{||s_i - s_j||^2}{2\sigma^2}\right)$ if $i \neq j$ and $A_{ii} = 0$
2. Define D to be diagonal matrix whose (i, i)–element is the sum of A's i–th row, and build the matrix $L = D^{-1/2} A D^{-1/2}$
3. Find $x_1, x_2, ..., x_k$ the k largest eigenvectors of L, and form the matrix $X = [x_1 x_2 ... x_k] \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns
4. Form the matrix Y from X by renormalizing each of X's rows to have unit length
5. Treating each row of Y as a point in $\mathbb{R}^k$, cluster them into k clusters
6. Finally, assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j

In step 5 we can apply k-means and then obtain an improvement of its straightforward implementation; avoiding convexity complications and running in a better computational way. The overall process is shown in Figure 1.

Summarizing, the top k eigenvectors[2] of the affinity matrix are used to form an $n \times k$ matrix Y where each column is normalized to unit length. Treating each row of this matrix as a data

[2]By the "top eigenvectors" we refer to the eigenvectors corresponding to the k smallest eigenvalues.

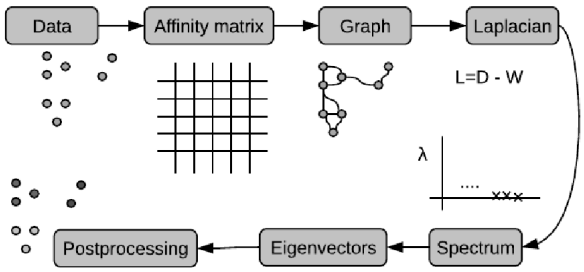

Figure 1: The process of Spectral Clustering: source Vejmelka [2009]

point, the algorithm of k-means is finally used to cluster the points. Now, our goal is to adapt this graphical process to the data information of a Water Supply Network. This can be done by building a kernel weight matrix to represent the WSN characteristics.

3.3 Semi–supervised clustering

Usually, when working on a real-world problem, some background knowledge about the cluster structure is available. We can assume that this knowledge comes in the form of pairwise must-link and cannot-link constraints. Such constraints are natural for graphs, as pairwise relationships are explicitly captured via edges in a graph (Kulis et al. [2005]). However, most semi–supervised clustering processes with pairwise constraints assume that the input is in the form of data vectors. This assumption is necessary in the problem of dividing the WSN into hydraulic zones. In this case, we should assure that each DMA is supplied by one or more water sources, such as tanks or reservoirs. This hydraulic feature should be translated as constraint implementation in the partition algorithm. In this work we show that a semi–supervised clustering solves this problem.

One spectral approach to semi–supervised clustering is the `spectral learning` algorithm of Kamvar et al. [2003]. A paper of Yu and Shi [2004] considered a semi–supervised formulation of the normalized cut objective and had a spectral algorithm associated with it. In this work we prefer the Kamvar solution to semi–supervised clustering, specifying penalty weights for constraint violations. As the entries of the affinity matrix are normalized between 0 and 1, we assign the next following penalty weights: $A_{ij} = 1 \; \forall i, j$ that have a must-link constraint and $A_{ij} = 0 \; \forall i, j$ that have a cannot-link constraint. At this stage, we can continue with the kernel spectral clustering algorithm, looking for the kernel associated to the affinity matrix and taking the top k eigenvectors to be the columns of the matrix to cluster.

3.4 The proposed algorithm

After the kernel data transformation (Section 2) we apply the proposed process to Water Supply clustering. To treat this data with the semi–supervised clustering methodology we first transform the graph affinity matrix into a kernel matrix. Next, we merge different information by adding other kernel matrices. This information contains values of the inputs under study, and the must-link and cannot-link constraints (see subsection 3.3). In this way, we make sure that each DMA is fed by at least one source (tank or reservoir). To perform the partition, we work with so many dissimilarity matrices (transformed into kernel matrices) as variables are involved. In the case that a previous DMA exists, it will be quantified in another matrix of distances. This matrix will be transformed to a kernel matrix and will be treated as an additional input (suitably weighted against all the other matrices). Then, the desired information is combined by using the weighted sum of kernel matrices with graphical and vector information.

This overall process (detailed in the current Section 3) can be summarized by the algorithm of the Table 2. We can add some improvements about this methodology working with the cluster

Table 2: Overall semi–supervised process

algorithm: water supply clusters by semi–supervised learning
1. abstraction of the water supply network as a graph
2. construction of Laplacian and dissimilarity matrices
3. data transformation into a single kernel matrix
4. *plug-in* of hydraulic constraints
5. calculus of the matrix spectrum
6. k-means into the top eigenvectors
7. cluster re-asignation into the original data
8. hydraulic validation (EPANET)

configuration quality. Some optimization of the weights in the kernel matrix construction (Table 2 – step 3) is other point to discuss in future researches.

4 EXPERIMENTAL STUDY

Starting from the affinity matrix of the graph of a WSN, the next step of the proposed methodology transforms the pipes, demand nodes and water constraints into a kernel matrix to which the development clustering algorithms are applied. In this Section we will apply the methodology exposed to the real case of the WSN of Celaya (Guanajuato, Mexico).

4.1 The Case Study

In order to show the performance of the presented process we consider here a real case, the WSN of the Central area of Celaya, fed by one reservoir ($D1$) and five tanks ($E1$, ..., $E5$) with five pump stations. This network is made out of 479 lines and 333 consumption nodes; its total pipe length is 42.5 km and the node elevation average is 156 meters; the total consumed flowrate amounts to 91 l/s.

In the case under study getting an optimal division of the WSN into DMAs is of paramount importance to improve the detection of leaks (more than 100 occurrences per year since 2005 to 2007) and to establish the necessary rehabilitation plans. The existing DMAs of the WSN will be taken into account as a starting point to run the algorithm proposed in this paper. This way, the information about the current sectorization, the analyses already performed and the built infrastructure may be used, instead of starting from the scratch.

4.2 Results

Our aim is to divide the WSN of our case study into 3 DMAs (each DMA being supplied by at least one tank). To this purpose, we apply the procedure of semi–supervised clustering algorithm explained in the present work[3]. After applying this process to the WSN data the following results are obtained. The size of each DMA in kilometers of pipes is 18, 9.5 and 15 km, respectively. The average diameter (in millimeters) of pipes per cluster is 144, 130 and 107, respectively. For this division, each DMA is supplied by at least one tank (see more details about average elevation and total demand of these hydraulic zones in Table 3). It is necessary to close 34 valves to isolate each DMA. The objective function of the clustering algorithm guarantees that it is the minimum number of links (valves) that should be closed for carrying out this graph partition.

In Figure 2 we can see, approximately, the cut plans dividing the WSN into three hydraulic zones. In addition, this figure shows the distribution of the different water suppliers in these DMAs. This final configuration was successfully simulated in EPANET, thus validating our results.

[3]Through an analysis of the cost we propose 0.4 as the weigth of the affinity matrix in the kernel matrix (eq.1). The remaining weights will be of equal importance to the sum of 0.6.

Table 3: Description of hydraulic zones of the case study

Sector	Nodes	Sources	Elevation	Demand
sector 1	122	E1 + D1	156.56	35.50
sector 2	84	E3 + E5	155.00	30.75
sector 3	127	E2 + E4	155.22	24.51

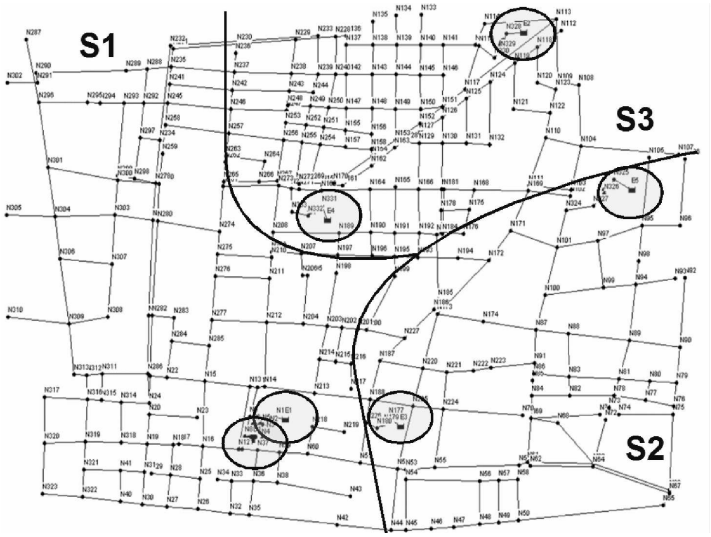

Figure 2: Aproximate scheme of the division of the WSN into three DMAs

5 CONCLUSIONS AND FUTURE RESEARCH

Classically, a division of a WSN into DMAs aims at improving leakage detection using node elevation, pressure and demand information. In the present work we propose augmenting, or changing, the perspective of this target. This can be done by taking into account different information to be included within criteria for the division of the WSN into hydraulic zones (clusters). Furthermore, one can use the diameter of the pipes and their age, weighting rehabilitation plans. Other point of view considers the use of some index of vulnerability for the pipes, taking into account the effects of hazards in the construction of DMAs.

Compared to other methodologies, which only use graphical or vector information, semi–supervised clustering use both, and in a more efficient and robust way. The flexibility to include different inputs into the study, with different weights, is another improvement of the shown methodology. Future research will focus on developing techniques to calibrate these weights. In addition, different modifications could be included in the clustering algorithm that could be compared on a work bench regarding their ability to build optimal models.

ACKNOWLEDGMENTS

This work has been carried out with the support of the project IDAWAS, DPI2009-11591, of the Dirección General de Investigación of the Ministerio de Educación y Ciencia (Spain) and with the collaboration of the Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS) of INSA, in Rouen (France).

REFERENCES

AVSA. Sectorización, 2009. Avalaible online http://www.aguasdevalencia.es/ last accessed in May 2009.

Covas, D. and H. Ramos. Practical methods for leakage control, detection and location in pressurised systems. In *13th International Conference on Pipeline Protection*, 1999.

Hunaidi, O. Economic comparison of periodic acoustic surveys and dma–based. In *Leakage 2005 Conference Proceedings*, pages 322–336, 2005.

Hunaidi, O. and K. Brothers. Optimum size of district metered areas. In *Water Loss Specialist Conference, International Water Association*, pages 57–66, 2007.

IWA. *District Metered Areas: Guidance Notes*. IWA Eds., 2007.

Izquierdo, J., M. Herrera, I. Montalvo, and R. Pérez. Agent–based division of water ddistribution systems into district metered areas. In *Proceedings of the 4th International Conference on Software and Data Technologies, ICSOFT 2009*, 2009.

Kamvar, S., D. Klein, and C. Manning. Spectral learning. In *In IJCAI*, pages 561–566, 2003.

Kondor, R. I. and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322. C. Sammut and A. Hofmann, editors, 2002.

Kulis, B., S. Basu, I. Dhillon, and R. Mooney. Semi–supervised graph clustering: a kernel approach. In *ICML'05: Proceedings of the 22nd international conference on Machine Learning*, pages 457–464. ACM, 2005.

Ng, A. Y., M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2001.

Rossman, L. *EPANET–User's Manual*. United States Environmental Protection Agency (EPA), 2000.

Scholkopf, B. and A. J. Smola. *Learning with kernels*. MIT Press, 2002.

Shi, J. and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

Smola, A. J. and R. I. Kondor. Kernels and regularization on graph. In *Proceedings of the 16th Annual Conference on Computational Learning Theory and the 7th Kernel Workshop*, 2003.

Tzatchkov, V., V. Alcocer-Yamanaka, and V. Bourguett-Ortíz. Graph theory based algorithms for water distribution network sectorization projects. In *8th Annual Water Distribution Systems Analysis Symposium*, 2006.

UKWIR. *A Manual of DMA Practice*. UKWIR Eds., London, 1999.

Vejmelka, M. Spectral graph clustering. In *Seminar z Umele Inteligence*, 2009.

Yu, S. X. and J. Shi. Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183, 2004.

Zhu, X., J. Kandola, J. Lafferty, and Z. Ghahramani. *Graph kernels by spectral transforms*. MIT Press, 2006.