CMN 2017

Congress on Numerical Methods in Engineering

July 3 - 5, Valencia, Spain

Edited by: Irene Arias, Jesús María Blanco, Stephane Clain, Paulo Flores, Paulo Lourenço, Juan José Ródenas and Manuel Tur

Congress on Numerical Methods in Engineering

CMN 2017

July 3 - 5
Valencia, Spain

Congreso de Métodos Numéricos en Ingeniería
3-5 julio 2017, Valencia, España
© SEMNI, 2017

MIXED COMPUTATIONAL AND HYDRAULIC CRITERIA FOR DMA CREATION USING HYBRID SOM+*k*-MEANS ALGORITHMS

Bernardo Novarini[1], Bruno Brentan[1], Gustavo Meirelles[1], Enrique Campbell[2], Edevar Luvizotto Jr.[1] and Joaquín Izquierdo[3]

1: Laboratory of Computational Hydraulics
School of Civil Engineering, Architecture and Urban Planning
University of Campinas
Av. Albert Einstein, 951, Cidade Universitária, Campinas, Brazil.
e-mail:{b.novarini@gmail.com, b080847@g.unicamp.br, limameirelles@gmail.com, edevar@fec.unicamp.br}
web: http://lhchidraulica.wixsite.com/grupoaguas

2: Berliner Wasserbetriebe
10864, Amtsgericht Charlottenburg, HRA 30951 B, Berlin, Germany
e-mail: enrique.campbellgonzales@bwb.de

3: FluIng - IMM
Universitat Politècnica de València
Camino de Vera, S/N (Edificio 5C-Bajo), 46022 Valencia, Spain
e-mail: jizquier@upv.es web: fluing.upv.es

Keywords: Water distribution system analysis, DMA creation, SOM, *k*-means.

Abstract *Integrated management of water supply systems with efficient use of resources requires optimization of operational performance. Following the "divide and conquer" strategy, clustering of water supply networks into small units, so-called district metered areas (DMAs), allows the development of specific operational rules, responsible for improving the network performance. In this context, clustering methods congregate neighboring nodes in groups according to some similar features, such as elevation or distance to the water source. Taking into account hydraulic, operational and mathematical criteria to determine the configuration of DMAs, this work presents a hybrid model SOM+k-means as a clustering model, comparing four quality-clustering indexes, namely Silhouette, GAP, Calinski-Harabasz and Davies-Bouldin, to determine the optimal number of clusters. Furthermore, to identify the best DMA configuration, the particle swarm optimization (PSO) method is applied to identify the number and location of DMA entrances. A large benchmark water network, EXNET, is used to evaluate the proposed methodology.*

1. INTRODUCTION

Water supply systems play a key role in urban design, not only to ensure that citizens have access to such an essential good, but also for public safety reasons [6,10]. The management of water supply systems has become increasingly complex in the face of the reduction of available natural resources, with the need to reduce energy consumption and water losses.

The division of the distribution network into district metered areas (DMAs) allows better management and increase of hydropower efficiency. However, such a division can be a complex task due to the size of the network and its peculiarities, such as the number of loops, the variation of the geometric dimensions and the modification in the hydraulic conditions, which may result in an inconsistent division if not considered [8].

Several studies have been proposed in the literature for the development of tools aiming at the automatic division of networks. [26] presented a model based on graph theory for the decomposition of water supply networks. [24] proposed the decomposition of multiple sources with the definition of pre-defined zones of influence for the network segregation. [14] used a multi-agent-based method to divide a water supply system into DMAs. [12] suggested a method for DMA creation based on machine learning methods. [8] proposed the automatic creation of boundaries for DMAs based on social structures, a tool in the field of Artificial Intelligence, and the theorem of decomposition of complex systems [21]. [4] made use of a clustering method based on social networks for the determination of DMAs, taking into account energy efficiency criteria. [7] proposed a method based on graph theory combined with an optimization algorithm for the determination of the DMAs, also targeting energy efficiency criteria. [2] presented a social community detection linked with a multi-level optimization to improve the management of water distribution systems.

Among the various clustering tools, the *k*-means algorithm is the most prominent. It was initially proposed by [22], and has been widely applied for clustering problems due to its simplicity, versatility and speed of operation [29], emphasizing its ability to handle a large amount of data [13]. Moreover, with the advent of modern neurology and the consequent discoveries of brain behavior, mathematical models of imitation of the behavior of this organ were proposed. Among them, [1,17,27] proposed the use of self-organizing maps that simulate the recognition of patterns by the brain for clustering, classifying, estimating and predicting various types of problems, with highlights to the study area of water resources.

However, the creation of DMAs in water supply networks is not fully achieved from data clustering. Once the DMAs have been defined, the inputs of each of the districts need to be determined, thus enabling the installation of control elements, such as pressure reducing valves that ensure the full insulation of a district in cases of emergency or maintenance. Current propositions make use of clustering and optimizing models to determine the sectors, and minimizing structural and deterioration costs [9].

Following current trends, this study aims to develop and analyze a model of DMA creation for water supply networks using mathematical criteria to define the optimal number of clusters. The model is based on a hybrid method, a combination of the self-organizing map

method (SOM) and *k*-means, with the purpose of determining the optimal cluster number of nodes with similar characteristics.
In a second stage, the challenge of finding optimal entrances for each DMA is solved by Particle Swarm Optimization (PSO). The minimization of costs, reaching the minimal pressure in the network, is conducted, resulting in the best scenario considering mathematical criteria and minimal costs.

2. METHODOLOGY

2.1. Self-organizing maps

Self-organizing maps have the main objective of processing input data in arbitrary dimensions and to bring them to one or two-dimension spaces through a transformation that guarantees topological similarity [11]. In general, the algorithm distributes a mesh of neurons within the feature space and along the iterations this mesh changes so that the synaptic weights are representative of the multidimensional space. The position, $\boldsymbol{w}_j$, of each node *j* of the network, also called a neuron, can be represented by:

$$\mathbf{w}_j = [w_{j1}, w_{j2} \dots w_{jm}]^T \; j = 1,2 \dots l, \tag{1}$$

where *l* is the total number of neurons in the network.
The search for the best similarity between a weight vector $\mathbf{w}_j$ and an input pattern $\mathbf{x}$ can be written as the minimization of a distance between both vectors. The neuron that satisfies the optimal condition is called the winning neuron and has associated to it a topological neighborhood that will define an activation zone, thus making a parallel to the biological inspiration behind the algorithm. The criterion of similarity is given by:

$$i(\mathbf{x}) = \text{argmin}_j\{||\mathbf{x} - \mathbf{w}_j||\}, \tag{2}$$

where $i(\mathbf{x})$ represents the winning neuron. The weights of the winning neuron and its neighboring neurons are then adjusted according to the following equation:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \mathrm{h}_C(t)[\mathbf{x}_i(t) - \mathbf{w}_j(t)], \tag{3}$$

where t represents the iteration of the training step, $\mathbf{x}_i(t)$ is the input pattern and $\mathrm{h}_C(t)$ is the neighborhood nucleus around the winning neuron.
The definition of the neighborhood usually follows the idea for which the activation of nearby neurons is greater than the activation of distant neurons. Figure 1 presents, in a simplified way, a self-organizing two-dimensional map with an *m* -dimensional input vector. The darker circle at the center represents the winning neuron and the gray scale shows the influence of the neighborhood in the adaptive process.

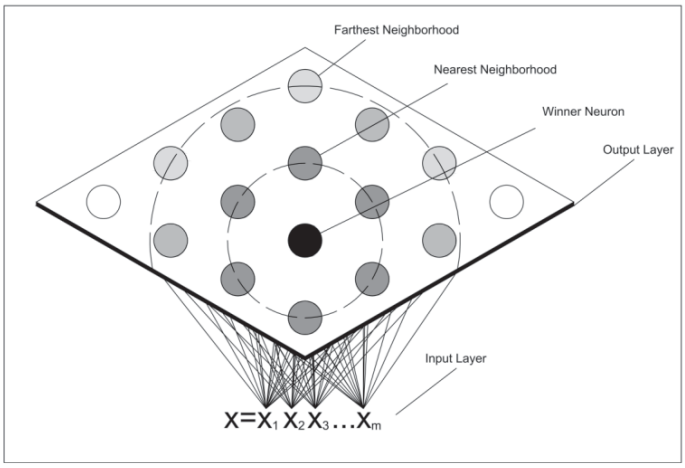

Figure 1: Example of a two-dimensional self-organizing map

Once the actuation neighborhood is defined, each of the weights is updated so that the topological proximity information is considered. With the learning process finalized, each neuron will be close to a certain set of input data represented in the output space. Each of the neurons can then be defined as the center of a cluster with a set of data around it.

2.2. *k*-means Algorithm

k-means is an unsupervised learning algorithm used to group the points of a network according to similar characteristics. The main point is the determination of centroids for hte clusters, being recommended to be located farthest from each other, and then allocate the points of the network to the nearest centroids. After such allocation it is necessary to recalculate the position of the centroids and evaluate if there was a change in their positions, repeating the process until no change occurs. An input vector $\mathbf{x}_i$ is represented in equation (4) and the objective function to be minimized by *k*-means is represented in equation (5).

$$\mathbf{x}_i = [x_1, x_2, x_3, x_4]^T \tag{4}$$

Here each component of vector $\mathbf{x}_i$ represents a feature of the data.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| \mathbf{x}_i^{(j)} - \mathbf{c}_j \right\|^2, \tag{5}$$

where $\left\| \mathbf{x}_i^{(j)} - \mathbf{c}_j \right\|^2$ is the Euclidian distance between a certain data $\mathbf{x}_i^{(j)}$ and the centroid $\mathbf{c}_j$, *k* is the number of centroids, and *n* is the number of nodes in the network.

2.3. Hybrid method

Clustering data with SOMs can result in large numbers of DMAs, which may not represent an optimal network partition. Therefore, the *k*-means algorithm is applied to achieve the ideal number of DMAs by reducing the number of clusters.

To take into account the previously clustered data from a SOM, *k*-means is applied to the final

position of each neuron. From the mathematical point of view, the final position of each neuron represents a set of data in the original feature space. From the computational point of view, the application of *k*-means to the neurons reduces the number of input data and, as a result, the process becomes fast. Finally, from the water distribution system analysis viewpoint, the reduction of the number of final clusters is important as it influences on how tractable de DMAs are.

In this sense, *k*-means receives the final position of each neuron and group the neurons in a pre-defined number of clusters. The optimal number of clusters is obtained through quality cluster analysis, using a number of mathematical criteria. The results represented by the centroid of each group are used to classify all data. Figure 2 shows the complete process of network partitioning.

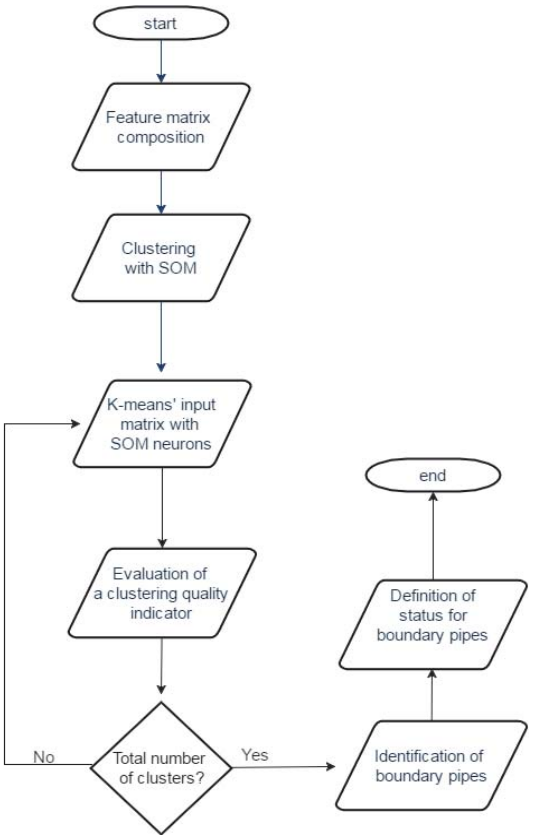

Figure 2: Flowchart of the clustering method and optimal entrance definition

2.4. Mathematical criteria for DMA creation

The main goal in clustering data is to determine groups with defined characteristics that differentiate as much as possible one from the others. The more compact the groups, the better the clustering tend to be, as this allows for less ambiguity. Thus, some measures of clustering quality are presented in the literature to evaluate the compactness of the clusters and the distance between them.

2.4.1. Gap

The Gap criterion [25] consists in obtaining a graph from the error measurements of the clustering procedure against the number of sectors of the network. The optimal number of sectors occurs when the greatest reduction of the related error occurs. A greater error reduction in relation to the number of sectors represents a higher gap value and, therefore, the optimal result occurs for the highest gap value, local or global, considering some tolerance limits.

Clustering the data into *k* clusters, the process to define the gap value starts by calculating the measure of dispersion within each sector:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{6}$$

where n_r is the number of data in cluster r, and D_r is the sum of the pairwise distances for all points in cluster r.

The gap value is defined as:

$$GAP_n(k) = E_n^*\{\log(W_k)\} - \log(W_k) \tag{7}$$

where n is the sample size and k is the number of clusters that maximizes $GAP_n(k)$, taking into account the sampling distribution. The expected value $E_n^*\{\log(W_k)\}$ is determined by the Monte Carlo method through a reference distribution and the $\log(W_k)$ is computed by the sample data.

2.4.2. Silhouette

The Silhouette criterion [15,20] is a value given by the analysis of similarity of a point in relation to the data of the same sector when compared to the data of other clusters. The silhouette value ranges from -1 to +1, where low or negative values represent poor results and high values represent appropriate sectoring results. This value is given by:

$$S_i = (b_i - a_i)/\max(a_i, b_i) \tag{8}$$

where a_i is the average distance of the i^{th} point in relation to other points in the same cluster and b_i is the smallest mean distance of the i^{th} point in relation to other points in different cluster.

2.4.3. Davies-Bouldin

The Davies-Bouldin criterion [5] is the ratio of distances within a given sector to distances between clusters. The Davies-Bouldin index is given by:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\{D_{i,j}\}, \tag{9}$$

where $D_{i,j}$ is the ratio of distances within the same cluster i and the distances between clusters i and j. In mathematical terms:

$$D_{i,j} = \frac{(\overline{di}+\overline{dj})}{d_{i,j}}, \quad (10)$$

where $\overline{di}$ is the mean distance between each point in the i^{th} cluster and its centroid, $\overline{dj}$ is the mean distance between each point in the j^{th}sector and its centroid, and $D_{i,j}$ is the Euclidean distance between the centroids of $i^{th}$and j^{th} sectors. The maximum value of $D_{i,j}$ represents the worst sector creation performed, while the minimum value represents optimal district creation.

2.4.4. Calinski-Harabasz (or VRC)

The Calinski-Harabasz criterion, or "variance ratio criterion" (VRC) [3], can be written as the relation between intra and inter-group distances, presented by equation (11),

$$VRC_k = \frac{SS_B}{SS_W} * \frac{(N-k)}{(k-1)}, \quad (11)$$

where SS_B is the total variance between clusters, SS_W is the total variance within each cluster, k is the number of clusters and N is the number of observations.

$$SS_B = \sum_{i=1}^{k} n_i \|\mathbf{m}_i - \mathbf{m}\|^2, \quad (12)$$

where k is the number of clusters, $\mathbf{m}_i$ is the centroid of sector i, $\mathbf{m}$ is the overall average of the sample data, and $\|\mathbf{m}_i - \mathbf{m}\|$ is the L^2-norm (Euclidean distance) between the two vectors.

$$SS_W = \sum_{i=1}^{k} \sum_{x \in c_i} \|\mathbf{x} - \mathbf{m}_i\|^2, \quad (13)$$

where x is a sample data, c_i is the i-th sector, $\mathbf{m}_i$ is the centroid of the cluster and $\|\mathbf{x} - \mathbf{m}_i\|$ is the Euclidean distance between the two vectors.
Well-defined sectors have high values of SS_B and low values of SS_W. The higher the VRC_k index, the better the sector creation for the data, as the optimum number of sectors is defined by the solution with the highest Calinski-Harabasz index.

2.5. Optimal entrances definition

The full isolation of a DMA requires the closure of a set of pipes that connect different DMAs and the installation of control and measurement devices, such as pressure reducing valves and flow meters. However, the closure of pipes modifies the hydraulic conditions and can harm the safe and continuous supply of water. Furthermore, the costs related with the closure of pipes and entrance devices are related to the diameter of the pipes where the devices will be installed. In this way, the choice of the set of closed pipes and entrances can be interpreted as an optimization problem, aiming to find the entrances with the lowest cost as possible while reaching the pressure limitation of the network.
The cost of the entrance devices C can be expressed by:

$$C = \sum_{i=1}^{n} s_i . c(D_i) \quad (14)$$

where n is the number of boundary pipes, s_i is the status of pipe i, that means if the pipe corresponding to an entrance of a DMA is open or closed, and $c(D_i)$ is the cost of the

entrance device installed on pipe i, with diameter D_i.
The optimal definition of entrances can be seen as the optimal definition of boundary pipes' statuses and the optimization problem can be written as:

$$\begin{gathered} \min(C) \\ s.t. \\ P_{min} < P_k \end{gathered} \tag{15}$$

where P_{min} is the minimal pressure required at the demand nodes, and P_k is the pressure at a node k.
The non-linearity of hydraulic equations turns hard the use of differential methods to solve the optimization problem. Particle Swarm Optimization [16] is one of the most common derivative-free, bio-inspired algorithms applied for water distribution network optimization [18,23].

3. RESULTS AND DISCUSSION

The methodology proposed in the paper was applied to the Exeter Network (EXNET) [28], composed of 1891 nodes, 3032 pipes, 7 water sources, with 2 major reservoirs, and 5 valves, with 3 check valves, 1 PRV and 1 throttle control valve. Figure 3 presents the topology of the EXNET. The hydraulic and topological features of the network are obtained by using Epanet2.0 [19] in MATLAB.

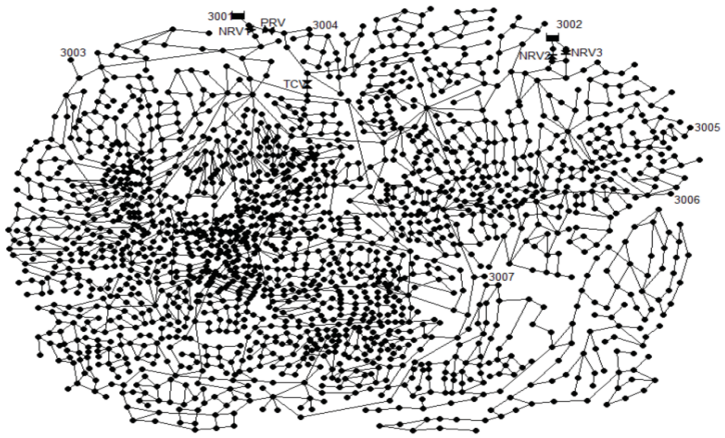

Figure 3: Arrangement of nodes and pipes of the Exeter Network displayed in EPANET 2.0 software

The SOM was set to have a grid dimension of 25 rows and 25 columns with a grid layer topology, running for a maximum number of 4000 iterations and an initial neighborhood size set to 4 neurons. This architecture resulted from a previous sensibility analysis, considering the processing time and the efficiency of the network, measured by the quantization error.

3.1. Mathematical criteria

Using the hybrid algorithm and considering the mathematical criteria, 4 network clustering were created. For each mathematical criterion, the quality of the district creation was evaluated using all the mathematical criteria, the quality indexes. Table 1 presents the index values for each of the mathematical criteria used, and Figures 5, 6 and 7 show the layout of the network for the mathematical criteria, with the Gap and Silhouette criterion presenting the same layout.

Table 1: Mathematical criteria index values for the quality evaluation of the DMA creation for each of the clustering criteria

Clustering Criterion	Quality Evaluation			
	Silhouette	Davies-Bouldin	Calinski-Harabasz	Gap
Silhoutte	0.60	0.95	1758.17	0.43
Davies-Bouldin	0.54	0.78	1713.11	0.34
Calinski-Harabasz	0.53	0.80	1759.14	0.33
Gap	0.60	0.95	1758.17	0.43

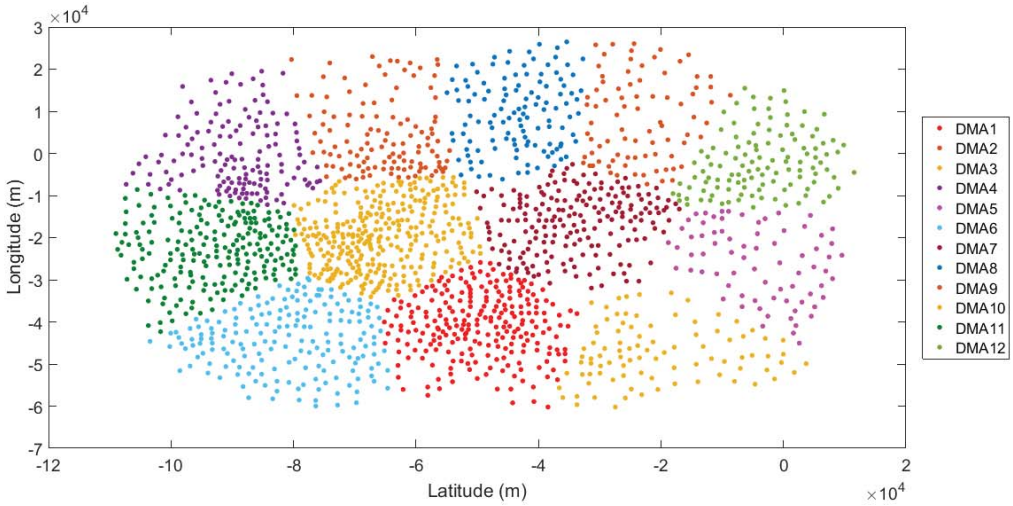

Figure 4: DMA creation in the EXNET using the Calinski-Harabasz mathematical criterion as the clustering criterion

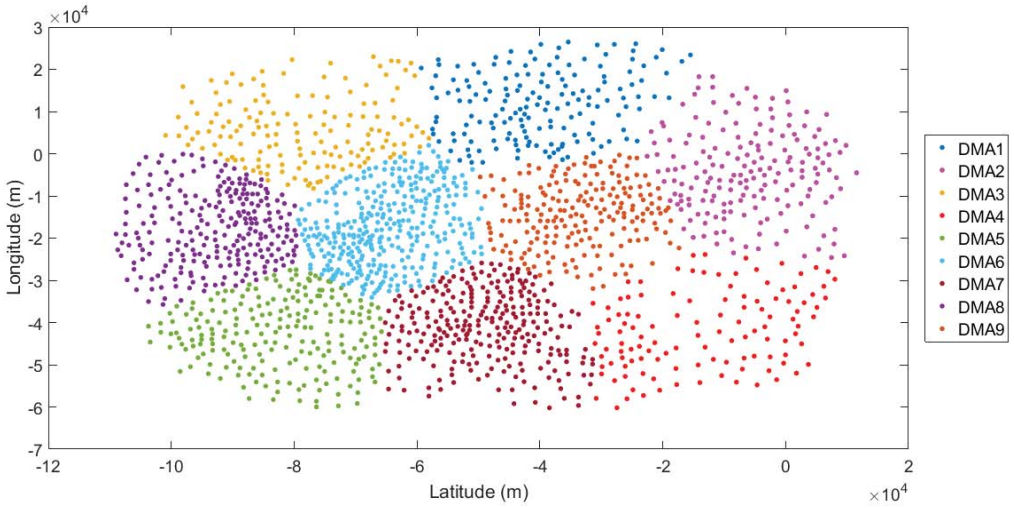

Figure 5: DMA creation in the EXNET using the Davies-Bouldin mathematical criterion as the clustering criterion

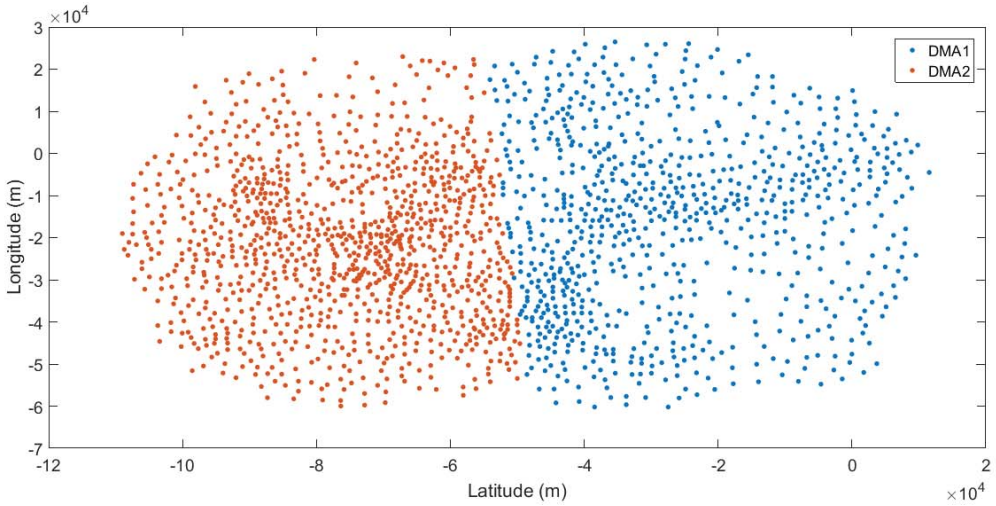

Figure 6: DMA creation in the EXNET using the Gap and Silhuette mathematical criteria as the clustering criterion

3.2. Optimization

The DMA creation for each of the mathematical criteria was optimized in order to analyze the cost involved for the optimal allocation of control valves and the insulation of the districts. Table 2 presents the results of the optimization for each of the mathematical criteria.

Table 2: Results of the optimization process for optimal entrance in terms of number of entrances and cost

Criterion	Number of DMAs	Number of boundary pipes	Number of control valves	Cost ($)	Unitary cost ($)
Silhouette	2	27	10	18500	1850.00
Davies-Bouldin	9	133	49	48787	995.65
Calinski-Harabasz	12	163	61	86255	1414.02
Gap	2	27	10	18500	1850.00

3.3. Discussion

It is possible to notice that DMA creation from the hybrid model presented well distributed and compact districts, representing a certain simplicity that facilitates the network management, which may result in an increase of overall efficiency of the water distribution system. The Silhouette and the Gap criterion presented bad results in the DMA creation, from the hydraulic viewpoint, with only 2 districts created, which does not represent a significant benefit to improve the management of the network. The Davies-Bouldin criterion presented a good result, with 9 compact well distributed districts throughout the network and good quality evaluation indexes. The Calinski-Harabasz criterion also presented a good result, with 12 compact well distributed districts throughout the network and the best quality evaluation index.

The total cost of entrance device installation represents an import investment for water utilities and increases with the number of DMAs. However, the highest unit cost is for the scenario with 2 DMAs (U$1,850) and the lowest for the scenario with 9 DMAs (U$995.65). This happens because as there is an increase in the number of DMAs, there are more boundary pipes and the possibility to find entrances with smaller diameters increases, representing a cost reduction.

Finally, considering the mathematical, hydraulic and implantation costs, although the Calinski-Harabasz criterion presented the best quality evaluation values, the scenario created by the Davies-Bouldin index can be looked as the most interesting one because it presents the lowest unit cost and good quality evaluation values. Furthermore, the partition of the network into 9 DMAs enables a more simplified operation and monitoring of the water distribution network.

4. CONCLUSIONS

The creation of DMAs in water distribution systems can be seen as a classical clustering problem and has been treated in various ways for several researchers. This work presented a hybrid model, SOM+*k*-means, considering the topological similarity of the nodes and several mathematical criteria to find the optimal number of clusters.

The topological similarity of the nodes in the water distribution network was essential for the effective creation of the DMAs. The hybrid model performed well, presenting good quality evaluation indexes and ability to simplify the water supply network, showing itself promising

for water distribution management.
Depending on the criterion used, the size and configuration of the DMAs will be unique, and it is up to the decision makers to choose the criteria that will be taken into account, considering the costs involved and the number of DMAs.
The use of mathematical criteria can generate impracticable solution from the hydraulic viewpoint. For future works, hydraulic criteria should be taken jointly with the mathematical criteria to improve the quality of DMA creation.

5. REFERENCES

[1] E. Alhoniemi, J. Hollmen, O. Simula and J. Vesanto. *Process monitoring and modeling using the self-organizing map*. Integrated Computer-Aided Engineering 6 (1), 3-14, (1999).

[2] B.M. Brentan, E. Campbell, G.L. Meirelles, E. Luvizotto and J. Izquierdo. *Social Network Community Detection for DMA Creation: Criteria Analysis through Multilevel Optimization. Mathematical Problems in Engineering*, (2017).

[3] T. Calinski and J. Harabasz. *A dendrite method for cluster analysis*. Communications in Statistics. Vol. 3, No. 1, pp. 1–27, (1974).

[4] E. Campbell, D. Ayala-Cabrera, J. Izquierdo, R. Pérez-García and M. Tavera. *Water supply network sectorization based on social networks community detection algorithms*. Procedia Engineering, 89, 1208-1215, (2014).

[5] D.L. Davies and D.W. Bouldin. *A Cluster Separation Measure*. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. PAMI-1, No. 2, pp. 224–227, (1979).

[6] A. Di Nardo and M. Di Natale. *A heuristic design support methodology based on graph theory for district metering of water supply networks*. Engineering Optimization, 43(2), 193-211, (2011).

[7] A. Di Nardo, M. Di Natale and G.F. Santonastaso. *A comparison between different techniques for water network sectorization*. Water Science and Technology: Water Supply, 14(6), 961-970, (2014).

[8] K. Diao, Y. Zhou and W. Rauch. *Automated creation of district metered area boundaries in water distribution systems*. Journal of Water Resources Planning and Management, 139(2), 184-190, (2012).

[9] E. Galdiero, F. De Paola, N. Fontana, M. Giugni and D. Savic. *Decision Support System for the optimal design of District Metered Areas*. J. Hydroinf. 18:1 49 – 61, (2015).

[10] W.M. Grayman, R. Murray and D. Savic. *Effects of redesign of water systems for security and water quality factors*. Proceedings of the World Environmental and Water Resources Congress, S. Starrett, (Eds.), May 17–21, Kansas City, Missouri, United States, 10.1061/41036(342)49, (2009).

[11] S. S. Haykin, *Neural networks: a comprehensive foundation.* Tsinghua University Press ,(2001).

[12] M. Herrera, S. Canu, A. Karatzoglou, R. Pérez-García and J. Izquierdo. *An approach to water supply clusters by semi-supervised learning.* Proceedings of International Environmental Modelling and Software Society (IEMSS), (2010).

[13] Z. Huang. *Extensions to the k-means algorithm for clustering large data sets with categorical values*. Data Min. Knowl. Discov. 2, 283-304, (1998).
[14] J. Izquierdo, M. Herrera, I. Montalvo, R. Pérez-García. *Division of Water Supply Systems into District Metered Areas using a Multi-agent based Approach. In: Software and Data Technologies (Communications in Computer and Information Science).* Springer-Verlag Berlin Heidelberg, 2011, Ch. 13, pp. 167-180, (2011).
[15] L. Kaufman and P.J. Rouseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, Inc., (1990).
[16] R.C. Eberhart and J. Kennedy. *A new optimizer using particle swarm theory.* In Proceedings of the sixth international symposium on micro machine and human science (Vol. 1, pp. 39-43), (1995).
[17] T. Kohonen. *Self-Organizing Maps.* Springer-Verlag, Berlin, (2001).
[18] I. Montalvo, J. Izquierdo, R. Pérez and M.M. Tung. *Particle swarm optimization applied to the design of water supply system*s. Computers & Mathematics with Applications, vol. 56, n. 3, pp 769-776, (2008).
[19] L.A. Rossman. *EPANET 2 USERS MANUAL*. U.S. Environmental Protection Agency, Washington, D.C., EPA/600/R-00/057, (2000).
[20] P.J. Rouseeuw. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics. Vol. 20, No. 1, 1987, pp. 53–65, (1987).
[21] H.A. Simon. *The Architecture of Complexity*. Proceedings of the American Philosophical Society Vol. 106, No. 6 (Dec. 12, 1962), pp. 467-482, (1962).
[22] H. Steinhaus. *Sur la division des corps matériels en parties.* Bull. I'acadèmie Pol. Sci. IV 801-804. C1. III, (1956).
[23] C.R. Suribabu and T.R. Neelakantan. *Design of water distribution networks using particle swarm optimization*. Urban Water Journal, vol. 3, n. 2, pp 111-120, (2006).
[24] Swamee and Sharma. *Design of Water Supply Pipe Networks*. John Wiley and Sons, (2008).
[25] R. Tibshirani, G. Walther and T. Hastie. *Estimating the number of clusters in a data set via the gap statistic.* Journal of the Royal Statistical Society: Series B. Vol. 63, Part 2, 2001, pp. 411–423, (2001).
[26] V.G. Tzatchkov, V.H. Alcocer-Yamanaka and V.B. Ortíz. *Graph theory based algorithms for water distribution network sectorization projects.* In Proc. of the 8th Annual Water Distribution Systems Analysis Symposium WDSA, Cincinnati, Ohio, USA, (2006).
[27] J. Vesanto and E. Alhoniemi. *Clustering of the self-organizing map*. IEEE Transactions on Neural Networks 11 (3), 586-600, (2001).
[28] Q. Wang, M. Guidolin, D. Savic and Z. Kapelan. *Two-Objective Design of Benchmark Problems of a Water Distribution System via MOEAs: Towards the Best-Known Approximation of the True Pareto Front. J. of Water Resources Planning and Management*. doi:10.1061/(ASCE)WR.1943-5452.0000460, (2014).

[29] X. Wu, V. Kumar, J.R. Quinlan, J. Gosh, Q. Yang, H. Motoda, G.J. McLachlan, A. NG, B. Liu, P.S. Yu, Z. Zhou, -H., M. Steinbach, D.J. Hand and D. Steinberg. *Top 10 algorithms in data mining*. Knowl. Inf. Syst. 14, 1-37, (2008).