RBRH

Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 24, e37, 2019
Scientific/Technical Article

https://doi.org/10.1590/2318-0331.241920180165

Optimal pressure management in water distribution networks through district metered area creation based on machine learning

Gerenciamento ótimo das pressões em redes de abastecimento de água através da criação de distritos de medição com base na aprendizagem de máquinas

Bernardo Novarini[1]; Bruno Melo Brentan[1]; Gustavo Meirelles[2] and Edevar Luvizotto Junior[1]

[1]Laboratório de Hidráulica Computacional, Faculdade de Engenharia Civil, Arquitetura e Urbanismo, Universidade Estadual de Campinas, Campinas, SP, Brasil
[2]Departamento de Engenharia Hidráulica e Recursos Hídricos, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
E-mails: b.novarini@gmail.com (BN), brunocivil08@gmail.com (BMB), gustavo.meirelles@ehr.ufmg.br (GM), edevar@fec.unicamp.br (ELJ)

Received: October 29, 2018 - Revised: May 22, 2019 - Accepted: June 10, 2019

ABSTRACT

Integrated management of water supply systems with efficient use of natural resources requires optimization of operational performances. Dividing the water supply networks into small units, so-called district metered areas (DMAs), is a strategy that 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 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 the k-means model and a hybrid model, that combines a self-organizing map (SOM) with the k-means algorithm, as clustering methods, comparing four mathematical criteria to determine the number of DMAs, namely Silhouette, GAP, Calinski-Harabasz and Davies Bouldin. The influence of three clustering topological criteria is evaluated: the water demand, node elevation and pipe length, in order to determine the optimal number of clusters. Furthermore, to identify the best DMA configuration, the particle swarm optimization (PSO) method was applied to determine the number, cost, pressure setting of Pressure Reducing Valves and location of DMA entrances.

Keywords: Water distribution network; Pressure management; Clustering; Optimization.

RESUMO

A gestão integrada dos sistemas de abastecimento de água com o uso eficiente dos recursos requer a otimização das operações. O agrupamento das redes de abastecimento de água em pequenas unidades, chamadas de distritos de medição (DMAs), é uma estratégia que permite o desenvolvimento de regras operacionais específicas, responsáveis por melhorar o desempenho da rede. Neste contexto, os métodos de classificação agrupam os nós vizinhos de acordo com características semelhantes, como elevação ou distância à fonte de água. Utilizando os critérios topológicos, operacionais e matemáticos para determinar a configuração dos DMAs, o trabalho apresenta um modelo k-means e um modelo híbrido, que combina um mapa auto-organizado (SOM) com o algoritmo k-means, como métodos de agrupamento. Comparou-se quatro critérios matemáticos, Silhouette, GAP, CalinskiHarabasz e Davies-Bouldin e analisou-se a influência de três critérios topológicos variáveis, a demanda de água, a elevação dos nós e o comprimento do tubo, para determinar o número ótimo de agrupamentos. Ademais, com o intuito de identificar a melhor configuração de DMAs, o método de otimização de enxame de partículas (PSO) foi aplicado para determinar o número, o custo, as pressões e a localização das entradas do DMA.

Palavras-chave: Rede de distribuição de água; Gerenciamento de pressões; Setorização; Otimização

INTRODUCTION

Water supply systems play a key role in urban design, not only to ensure that citizens can have access to essential goods, but also for public safety reasons (DI NARDO, DI NATALE, 2011; GRAYMAN et al., 2009). The management of water supply systems become increasingly complex in the face of the reduction of available natural resources, with the need to reduce energy consumption and water loss.

The division of the water distribution network (WDN) into districts allows a better management and increase of hydraulic and energy efficiency, since the operations are directed to the needs of each district, besides the greater control from measurements and monitoring. However, such 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 can make such a division inconsistent if they are not considered (DIAO et al., 2012).

For the definition of a district metered area (DMA) it is necessary to determine the supply points (entrance points) and their influence regions. In this definition, water supply should provide sufficient quantity and quality to consumers. Operating pressures must be ensured inside a standardized range, a condition normally achieved by using pressure reducing valves (PRVs). The location of supply points in the district and the operating pressure are fundamental in the clustering process.

Corroborating the importance of the division of networks into districts, important work are proposed in the literature for the development of clustering tools. Tzatchkov et al. (2006) present a model based on graph theory for the segmentation of supply networks. The authors were based on graph analysis and graph partition in order to find a suitable design for the DMAs. Swamee and Sharma (2008) propose the segmentation of multiple sources assigning pre-defined zones of influence for the clustering. Herrera et al. (2010) proposed the use of partitioning with methods based on machine learning for the definition of DMAs. Also based on graph partition, the authors included the non-supervised learning approach to the DMA design, developing a hybrid graph theory / data mining algorithm for the DMA design. Diao et al. (2012) proposed the automatic creation of boundaries for the determination of measurement districts based on social structures, a tool in the field of Artificial Intelligence, and the decomposition theorem of complex systems (Simon, 1962). Campbell et al. (2014) proposed a clustering method based on social networks for the determination of districts using energy efficiency as criteria. In this work, the authors found a robust and computational efficient technique for DMA design in large networks based on graph partitioning and data mining technique for nodal clustering. The authors used topological criteria, such as the maximal demand of a district, or the maximal difference in node elevation as criteria for the graph partition. Di Nardo et al. (2014) proposed a method based on graph theory coupled to an optimization algorithm for the determination of the districts of a supply network also aiming energy efficiency improvement.

Among the several clustering tools, the k-means algorithm is the most prominent. Initially proposed by Steinhaus (1956), it is widely used for clustering problems due to its simplicity, versatility and speed of operation (WU et al., 2008), emphasizing its ability to handle a large amount of data (HUANG, 1998). On the other hand, with the advent of modern neurology and the consequent discoveries of cerebral functioning, mathematical models based on the behavior of this organ were proposed. Among them, Alhoniemi et al. (1999), Vesanto and Alhoniemi (2000) and Kohonen (2001) proposed the use of a self-organizing map (SOM), which simulates the recognition of patterns by the brain for grouping, classifying, estimating and predicting different types of problems, being widely used in the area of water resources.

The challenge of creating DMAs in supply networks is not fully solved from a database. Once defined the districts, it is necessary to define the entrance of each of these districts, thus allowing the installation of control elements, such as PRVs, to ensure complete isolation in cases of emergency or maintenance. The current propositions make use of hybrid optimizer-cluster models to determine the districts, minimizing structural costs and deterioration (GALDIERO et al., 2015).

During the last decades, the water companies have developing to divide the water network, aiming a better management. The recommendation of United Kingdom, early of 1980's (FARLEY, 1985) has change the management of water distribution systems and, by the strategical placement of pressure control devices, the leakage rate could be reduced. Nevertheless, the task to create DMAs is still a complex task because many variables are playing important rules, such as topological and topographic features, costs, benefits etc. In order to develop an automatic tool for DMA design coupled to the optimal pressure management, this work develops and analyzes two models of DMA creation in water supply networks using two sets of criteria, the mathematical and topological. The first model is based on the k-means clustering algorithm and the second one is a hybrid method, combining the SOM and k-means methods, both with the purpose of determining the optimum number of groups of nodes with similar characteristics. Four mathematical criteria to determine the number of DMAs are evaluated, namely Silhouette, GAP, Calinski-Harabasz and Davies Bouldin. In addition, the influence of three clustering topological criteria is evaluated: the maximal water demand, maximal difference in node elevation and total pipe length. Finally, an optimization model, based on the bio-inspired particle swarm optimization (PSO) algorithm, is applied for the allocation of control and isolation valves of the districts, as well as their operation point, minimizing the installation costs.

In this sense, the purposed method is composed by 2 stages. The first one, based on physical (elevation) and topological (space position) parameters of the networks, a clustering algorithm (K-means) is applied. The algorithm will divide the network in K groups, based on Euclidian distances from K-centers, initially randomly distributed, and recurrently self-organized, based on the mean value of each k-group. The important task at this stage is to define the value of K. To help the solution of this task, mathematical and topological criteria are explored in this paper. Each criterion is considered separately. For future works, mainly for the topological criteria, the analysis of correlation or interference between the criteria could be considered.

MATERIAL AND METHODS

Clustering algorithms

Self-organizing maps

The main objective of a SOM is to process input data in arbitrary dimensions and bring them to a one or two-dimensional set of data, with transformations that guarantees topological similarity (HAYKIN, 2001). In general, the algorithm distributes a group of neurons within the characteristic space and as iterations occur, this group changes so that the synaptic weights are representative of the multidimensional space, without previous knowledge of the behavior of such surface.

The position of each node *j* of the network w_j, also called the neuron can be represented by equation 1:

$$w_{ji} = \left[w_{j1}, w_{j2} \ldots w_{jm} \right]^T \text{ j=1,2... N;} \quad (1)$$

where N is the total number of neurons in the network.

The similarity between a weight vector w_{ji} and an input pattern x_i can be measured in terms of the distance between the two vectors. The neuron that satisfies the optimal condition of minimum distance is called the winning neuron and has associated to it a topological neighborhood that will define an activation zone. The criterion of similarity is given by equation 2:

$$\|x - w_C\| = min_j \left\{ \|x - w_j\| \right\} \quad (2)$$

in which $\|x - w_C\|$ represents the Euclidean distance between the network neurons and c represents the chosen winning neuron.

The weights of the winning neuron and its neighboring neurons are then adjusted according to the following equation 3:

$$w_{ji}(t+1) = w_{ji}(t) + h_C(t)\left[x_i(t) - w_{ji}(t) \right] \quad (3)$$

where t represents the iteration of the training, $x_i(t)$ is the input pattern and $h_C(t)$ is the neighborhood nucleus around the winning neuron.

The definition of the neighborhood usually follows the idea in which the activation of nearby neurons is greater than the activation of distant neurons. Figure 1 presents, in a simplified way, a two-dimensional SOM with a two-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.

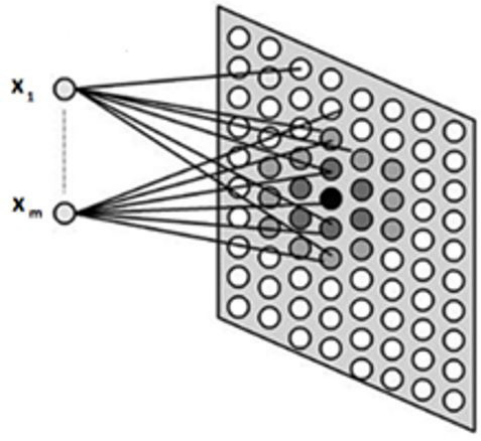

Figure 1. Example of a two-dimensional self-organizing map (adapted from Koua and Kraak (2006)).

Once the actuation neighborhood is defined, each of the weights is updated so that all 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, then labeled.

k-means

K-means is an unsupervised learning algorithm used to group the points of a network according to similar characteristics. The algorithm works by determining the centroid for each cluster. The best clustered data will have their centroids located farthest from each other, allocating the points of the network to the nearest centroids. The *k* centroids are selected randomly in the input space and each input data is classified according to their distance to the centroids. After the allocation, it is necessary to recalculate the position of the centroids and evaluate if there is any change regarding the previous position, repeating the process until there are no changes. The new position of the centroids is calculated with equation 4.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} x_i^{(j)} - c_j^2 \quad (4)$$

where $x_i^{(j)} - c_j^2$ is the distance between an input vector $x_i^{(j)}$ and the centroid c_j, *k* is the number of centroids and *n* is the number of nodes in the network. In this study, the input vector x_i has four dimensions, representing the demand, elevation, latitude and longitude of node *i* of the network, as shown in equation 5.

$$x_i = \left[x_1, x_2, x_3, x_4 \right]^T \quad (5)$$

Criteria for clustering in districts

Clustering criteria are used to feed the algorithms with information in order to identify similar network nodes, grouping them in specific DMAs. Two types of criteria were considered for clustering: topological and mathematical. The first takes into consideration only the physical features of water supply networks. The second considers the quality of the clusters created.

Topological criteria

The topological criteria of a water supply system such as the maximal water demand, maximal difference in node elevation and total pipe length. define the hydraulic behavior of the network. Identifying such criteria in the clustering process can favor the pressure management in the districts.

The maximum water demand, the maximum elevation difference between nodes and the maximum pipe length of the

same district were used to determine the number of clusters, varying the limit values of each one separately to verify the influence of each factor.

Mathematical criteria

The main purpose of clustering data is to determine groups with solid characteristics that differ as much as possible from each other. In addition, the more compact the clusters, the less ambiguity the overall clustering. Thus, measurements of quality are shown in the literature as means to evaluate both the distance between clusters and their compactness. The mathematical criteria used for the quality-cluster analysis were: GAP, Silhouette, Davies-Bouldin and Calinski-Harabasz.

GAP

The GAP criterion (TIBSHIRANI et al., 2001) consists of obtaining a graph of error measurements in the clustering relating to the number of clusters of the network. The optimal clustering occurs when the maximum reduction of related error is achieved. Reduction in errors in relation to the number of clusters represents higher GAP values, with the optimal result occurring at the highest GAP value, local or global, considering tolerance limits. The GAP value is defined as shown in equation 6:

$$GAPn(k) = E^{*}n\{log(Wk)\} - log(Wk) \tag{6}$$

where n is the sample size, k is the number of clusters being evaluated and Wk is the measure of dispersion within each cluster. The expected value $E^{*}n\{\log(Wk)\}$ is determined by the Monte Carlo method through a reference distribution and the log (Wk) is computed by the sample data, as shown in equation 7.

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} * D_r \tag{7}$$

where n_r is the number of data in a cluster r, and Dr is the sum of the distance between all points of cluster r.

Silhouette

The Silhouette criterion (ROUSEEUW, 1987; KAUFMAN; ROUSEEUW, 1990) consists of a similarity analysis of specific data points in relation to the data of the same cluster compared in relation to the data of other clusters. The silhouette value ranges from -1 to +1, with low or negative values representing poor results and high values representing appropriate clustering results. This value is given by equation 8:

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

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 clusters.

Davies-Bouldin

The Davies-Bouldin criterion (DAVIES; BOULDIN, 1979) consists of a ratio of the distance of nodes within a given cluster to the distance between clusters. The Davies-Bouldin index is given by equation 9:

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

where $D_{i,j}$ is the ratio of distances within the same cluster i and the distances between clusters i and j. The equation 10 shows the ratio of distance in mathematical terms:

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

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} cluster and its centroid, and $D_{i,j}$ is the Euclidean distance between the centroids of i^{th} and j^{th} clusters. The maximum value of $D_{i,j}$ results in the worst DMA creation performance, while the minimum value represents optimal creation.

Calinski-Harabasz

The Calinski-Harabasz criterion, or "variance ratio criterion" (VRC) (CALINSKI; HARABASZ, 1974), consists of the relation between intra-cluster distances. The VRC is given by equation 11:

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

where SS_B is the total variance between clusters - equation 12, SS_W is the total variance within each cluster equation 13, k is the number of clusters and N is the number of observations.

$$SS_B = \sum_{i=1}^{k} n_i m_i - m^2 \tag{12}$$

where m_i is the centroid of cluster i, m is the overall average of the sample data, and $m_i - m$ is the L^2 norm (Euclidean distance) between the two vectors.

$$SS_W = \sum_{i=1}^{k} \sum_{x \in c_i} x - m_i^{\,2} \tag{13}$$

where x is a sample data, c_i is the i^{th} cluster and $x - m_i$ is the L^2 norm (Euclidean distance) between the two vectors.

High values of SS_B and low values of SS_W represent well defined clusters. The higher the VRC_k index, the better the clustering, with optimum number of clusters defined by the solution with highest Calinski-Harabasz index.

Optimal pressure management

Considering the optimum pressure management within each of the DMA, it is proposed in this study the optimal allocation of valves in the entrance of each district, and their pressure setting, aiming the highest uniformity of pressure within the district.

The choice of the nodes belonging to a previously grouped DMA should comply with the minimum and maximum pressure constraints in addition to operational criteria that are raised throughout the study and enable better management of the districts.

Considering as decision variables the location of each of the valves and their respective pressure setting, the problem can be written as the minimization of the operating pressures of the system and the pressure uniformity parameter (PU_k), which expresses the pressure deviation of each node with respect to the mean pressure of the nodes of a district. This measure was proposed by Alhimiary and Alsuhaily (2007) shown in equation 14. The minimization problem is subject to the pressure constrains (Equation 15) and the number of nodes belonging to a DMA (equation 16).

$$Min\ PU_k = \sum_{t=1}^{T}\sum_{i=1}^{N_k}\left(\frac{\sqrt{\frac{\left(P_{i,t}-P_{m_k,t}\right)^2}{N_k}}}{P_{m_k,t}}\right) \tag{14}$$

$$P_{min} \le P_{i,t} \le P_{max} \tag{15}$$

$$N_k \le N_{máx} + \varepsilon \tag{16}$$

where PU_k is the pressure uniformity parameter for a given district k, T is the total simulation period, N_k is the number of nodes belonging to district k, $P_{i,t}$ is the pressure at a given node i for the time step t, $P_{mk,t}$ is the mean pressure of district k in time step t, P_{min} and P_{max} are the minimum and maximum standardized pressures respectively. The bio-inspired Particle Swarm Optimization (PSO) algorithm is used to determine the position of the valves and respective pressure settings.

Particle Swarm Optimization - PSO

Particle Swarm Optimization (PSO) is a population-based algorithm that has particles as the elemental unit. The particles are composed of two vectors of size D (dimension of the problem). One of these vectors represents the position of the particle and the other its displacement velocity. The first step of the method is the initialization of the particles, done randomly within a range of interest, both for position and for the velocity. At each iteration n, the particle information is updated, considering its best position ever achieved (p_{id}) and the group best position (g_{id}) as shown in Equations 18 and 19 (EBERHART; KENNEDY, 1995). The process continues until one of the stopping criteria is reached, such as the maximum value with the arbitrated error, the maximum number of iterations, the lack of improvement in the objective function for a determined iteration interval and other stopping criteria widely used in numerical problems (FAIRES; BURDEN, 1998)

$$v_{id}^{n+1} = \left[w * v_{id}^{n} + \frac{c_1 * r_1 * \left(p_{id}^{n} - x_{id}^{n}\right)}{\Delta t} + \frac{c_2 * r_2 * \left(g_{id}^{n} - x_{id}^{n}\right)}{\Delta t} \right] \tag{18}$$

$$x_{id}^{n+1} = x_{id}^{n} + v_{id}^{n+1} * \Delta t \tag{19}$$

where $d = 1,2, \ldots m$, with m the number of variables of the problem, $n = 1,2, \ldots N$, with N the maximum number of iterations. Also, r_1 and r_2 are numbers randomly chosen within the range [0,1], and c_1 and c_2 the cognitive and social coefficients respectively. The first is used in the initial iterations to perform a global search, while the second improves the local search, for the final iterations, when it is expected to be close to an optimal solution.

RESULTS AND DISCUSSION

The method proposed was applied to the D-Town network (MARCHI et al., 2013), composed of 398 nodes, 458 pipes, 7 tanks, 1 reservoir, 13 pumps and 4 valves, as shown in Figure 2.

The SOM was configured to have 25 rows and 25 columns with a squared topology, to execute a maximum number of 4000 iterations and a defined topological neighborhood size of 4 neurons. This arrangement was chosen through a sensitivity analysis, considering the processing time and the efficiency of the algorithm, measured by the quantification of the errors.

Figure 2. Arrangement of the nodes and pipes of the D-Town network made in EPANET 2.0 software.

Topological criteria

A total of 18 scenarios were generated, 9 with the k-means algorithm and 9 with the hybrid algorithm, varying the district's maximal water demand, maximal difference in node elevation and total pipe length for the district. For each criterion, the cluster quality was evaluated using the Calinski-Harabasz index (VRC), in which a higher index value represents a higher quality of DMA creation.

Starting with the demand criterion, Table 1 presents the VRC index values for each of the limits used. There is a slight difference between the demand limit of 140 l/s when compared to the other values for hybrid clustering. Still, the best value of VRC is obtained by creating DMAs with the k-means method. Figure 3 presents the best creation scenario for each of the methods using the demand of 140 l/s as the limit value.

It can be noticed a spatial difference of the clustering patterns between one method and another. The k-means method generates more circular districts, around a center of gravity, which is more compatible with reality.

Following the evaluation of the criterions for DMA creation, Table 2 shows the value of the VRC index using the elevation as parameter. In both methods the best value for VRC occurs with the maximum elevation difference of 75 m. Figure 4 shows the final distribution of the districts for each of the algorithms.

The last topological criterion analyzed was the maximum total pipe length for the district. Table 3 shows the value of the VRC index for each of the criteria boundaries. It is observed that the district with a maximum of 15 km has the best performance, and the clustered network for this limit value, in each one of the algorithms is presented in Figure 5.

Table 1. Calinski-Harabasz index for the variation of the district water demand criterion for models with k-means and hybrid algorithms.

Demand	k-means	Hybrid
100 l/s	408.05	414.63
120 l/s	417.17	414.63
140 l/s	417.43	416.58

Table 2. Calinski-Harabasz index for the variation of the criterion of maximum node elevation difference between districts for the models with k-means and hybrid algorithms.

Elevation	k-means	Hybrid
75 m	417.54	362.05
80 m	416.62	280.27
90 m	362.05	280.27

Table 3. Calinski-Harabasz index for the variation of the maximum total pipe length for the districts regarding the models with k-means and hybrid algorithms.

Pipe Length	k-means	Hybrid
15 km	417.59	280.27
20 km	416.62	280.27
25 km	362.05	280.27

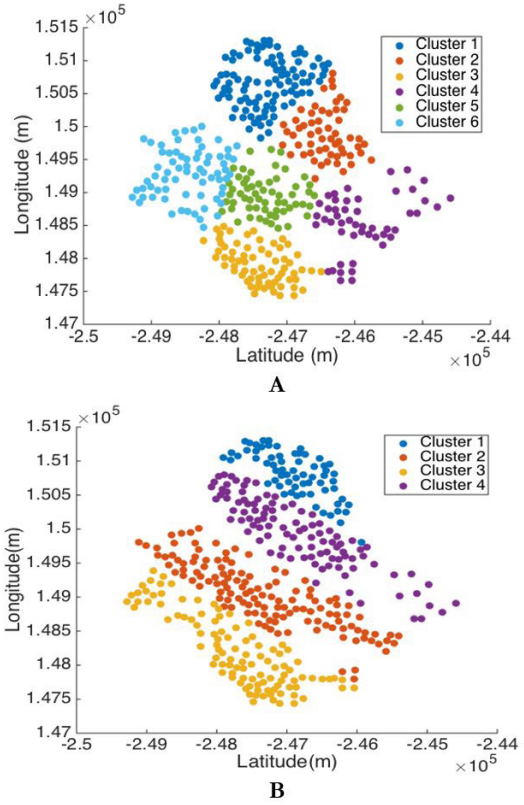

Figure 3. Clustering of the D-Town network with the k-means algorithm (A) and hybrid algorithm (B) with the maximum district water demand of 140 l/s.

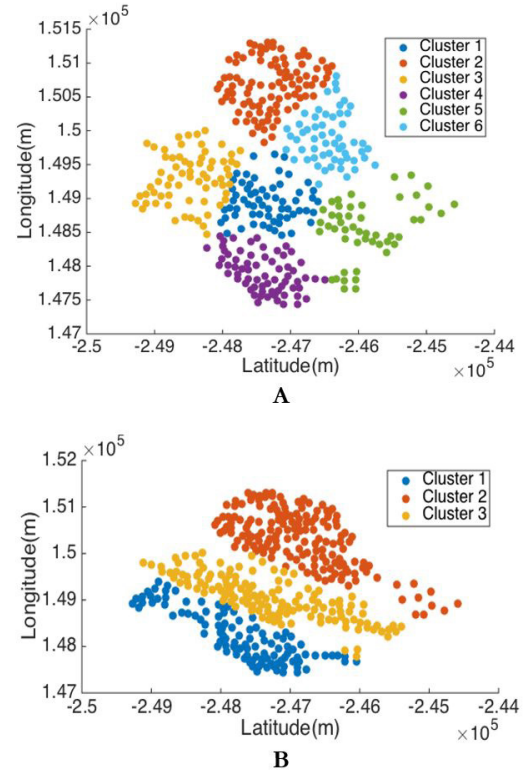

Figure 4. Clustering of the D-Town network with the k-means algorithm (A) and hybrid algorithm (B) with the maximum node elevation difference between districts of 75m.

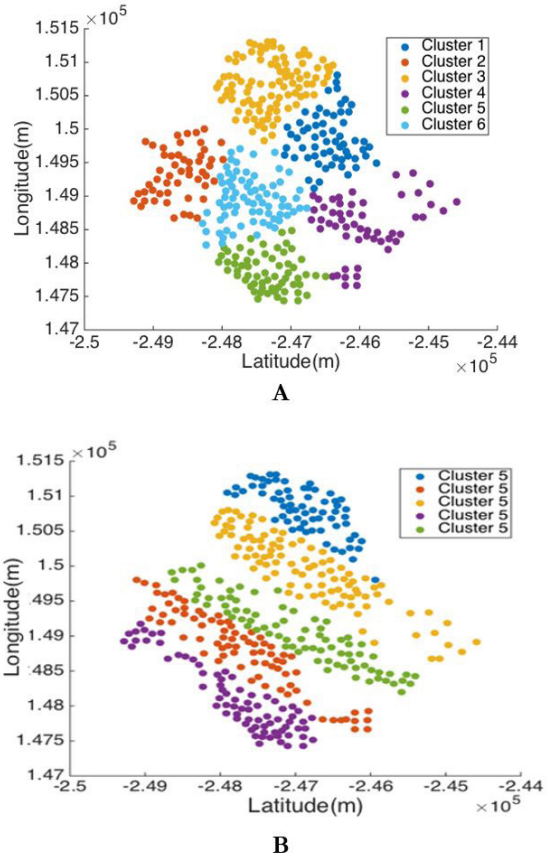

Figure 5. Clustering of the D-Town network with the k-means algorithm (A) and hybrid algorithm (B) with maximum total pipe length for the district of 15km.

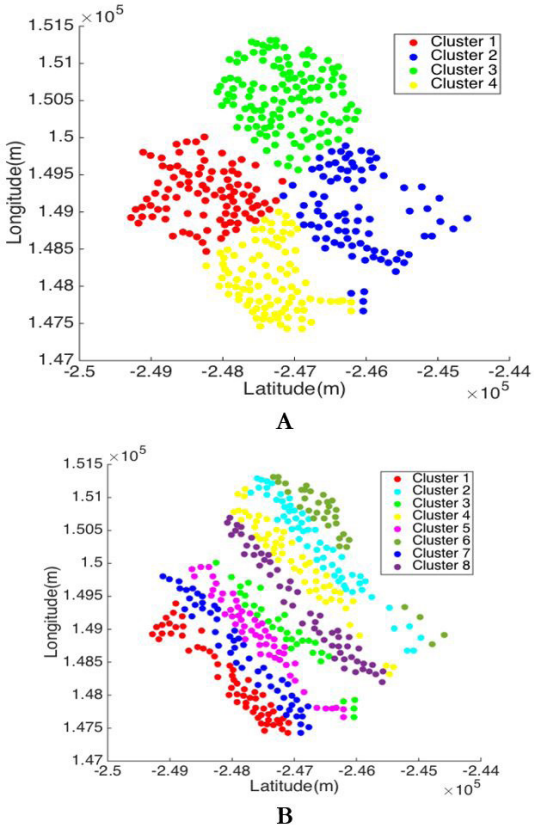

Figure 6. Clustering of the D-Town network with the k-means algorithm (A) for the scenario generated by the VRC criterion and the hybrid algorithm (B) for the scenario generated by the DB criterion.

Within the topological criteria, the one that presented the best performance, when evaluated by the VRC index, was the scenario generated by the k-means algorithm with the maximum district length criterion. This result is very close to the districts generated by the same algorithm with the maximum demand criterion. In general, the k-means algorithm presented better performance alone when compared to the districts generated by the hybrid model.

Mathematical criteria

A total of 8 scenarios were generated, 4 with the k-means algorithm and 4 with the hybrid algorithm, varying the mathematical criteria. For each mathematical criterion, the quality of the district was also evaluated using the Calinski-Harabasz index. Table 4 shows the value of the VRC index for each of the mathematical criteria used.

It can be noticed, for the k-means method, the scenario obtained by the VRC criterion itself had the best result, similar with those found in the topological criteria. On the other hand, the evaluation of the hybrid model had a better result with the scenario generated by the Davies-Bouldin criterion (DB), but once again, in all cases of the hybrid model, the clustering had lower quality values than the method k-means pure. Figure 6 shows the final distribution of the districts for each of the criteria.

Table 4. Calinski-Harabasz index for the variation of the mathematical criterion for models with k-means and hybrid algorithms.

Criterion	k-means	Hybrid
GAP	362.04	409.13
Silhouette	416.58	280.27
Davies-Bouldin	416.71	416.58
Calinski-Harabasz	417.36	408.95

Optimization of entrance location and operational point of VRPs

For each criterion used in the creation of DMAs, an optimization was performed on the k-means method, with the purpose of analyzing the cost involved in the optimal allocation of PRVs and the distribution of the pressures in the network under conditions of maximum and minimum demand for a period of 24 hours. The choice of k-means models is justified because they presented better results in the creation of DMAs, with well-distributed and compact clusters.

The total cost represents the cost involved in the installation of PRVs, while the unit cost represents the cost per valve implanted. The cost for PRV are based on Saldarriaga et al. (2019). This analysis was made in order to obtain insights on the costs associated with the pressure optimization. Table 5 presents the optimization results for each criterion.

A good pressure distribution in the network occurs when the operating pressures of the system and the standard deviation between them are minimized, both in the conditions of minimum and maximum demand, comparing to the situation without optimization. The topological criteria presented an improvement in the distribution of pressure in the network, with emphasis on the "Length 15 km" criterion, which showed a significant reduction in the pressure required by the network, evident in Figure 7. The mathematical criteria also showed an improvement in the distribution of pressure in the network, with Calinski-Harabasz, which presented a significant reduction in the pressure required by the network, evident in Figure 8.

DISCUSSION

It is possible to notice from Figures 3-6 that the models that only used k-means to group the nodes of the network had a well distributed and compact aspect districts. By using the hybrid model, it is possible to notice that all the clusters maintained the same pattern of clustering in diagonal bands, losing the essence of compact clusters and possibly representing difficulties in the strategic management of the districts, since they have an elongated aspect.

The variation of the topological criteria resulted in changes in the arrangements and number of districts, in which the increase of the criteria values tended to reduce the number of districts.

Table 5. Results of optimization of control valve allocation for each criterion used to create the DMAs.

	Criterion	District Metered Area (DMA)	Closed Pipes	#Control Valves	Cost ($)	Unit Cost ($)
Qual.	Silhouette	2	8	7	13,456	1,922
	Davies-Bouldin	2	9	8	16,435	2,054
	Calinski-Harabasz	8	81	55	97,886	1,780
	GAP	2	9	8	16,435	2,054
Topol.	Demand	6	24	17	33,092	1,947
	Elevation	6	24	18	40,492	2,250
	Length	6	22	17	42,716	2,513

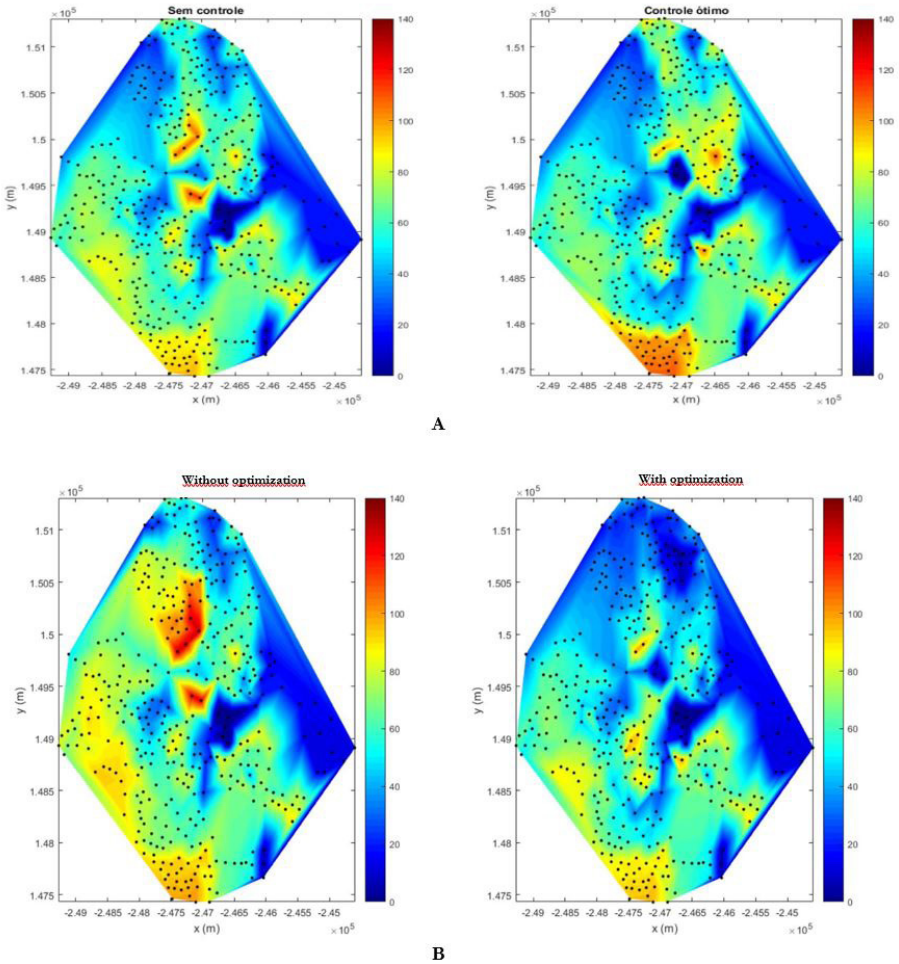

Figure 7. Distribution of the pressures in the D-Town network, without and with optimization, for the mathematical criterion "Calinski-Harabasz" in conditions of maximum (A) and minimum demand (B).

The mathematical criteria did not show drastic differences among them, with the Calinski-Harabasz criterion presenting the largest number of districts and the GAP criterion the lowest number of districts in the case of the model using only k-means.

When analyzing the Calinski-Harabasz index in the clustering, it is possible to notice that the models with the k-means algorithm presented, in general, higher indexes, thus with a higher quality. The best clustering with respect to the topological criteria was given for the maximum DMA water demand equal to 140 L/s (6 DMAs generated), the difference in node elevation between DMAs equal to 75 m (6 DMAs generated) and the maximum total pipe length of the DMA equal to 15 km (6 DMAs generated). The best clustering in relation to the mathematical criteria was obtained by using the Calinski-Harabasz method (8 DMAs generated), although the methods Silhouette (2 DMAs generated) and Davies-Bouldin (2 DMAs generated) presented very close indexes.

When analyzing the creation of DMAs in terms of mathematical criteria, the Silhouette, Davies-Bouldin, and GAP presented poor hydraulic results with only 2 DMAs created, which is not a significant improvement for management purpose. The Calinski-Harabasz criterion presented a good result, with 8 compact districts well distributed throughout the network and good quality evaluation indexes, in addition to a lower unit cost for PRVs installation (U$ 1,780).

When analyzing the creation of DMAs in terms of the topological criteria, all presented good results, with 6 DMAs created, compact and with well distributed characteristics. The criterion "Demand 140 l/s" presented the lowest total cost (U$ 33,092) and unit cost (U$ 1,947) for PRVs installation.

It is possible to notice that the total cost of installation increases with the number of DMAs. However, the unit cost tends to decrease, as there are more limit tubes and more likely to work with smaller diameters, reducing the costs of PRVs

Figures 7 and 8 highlight the efficiency of the network optimization as to the distribution of pressures under the conditions of minimum and maximum demand of the system, reducing the

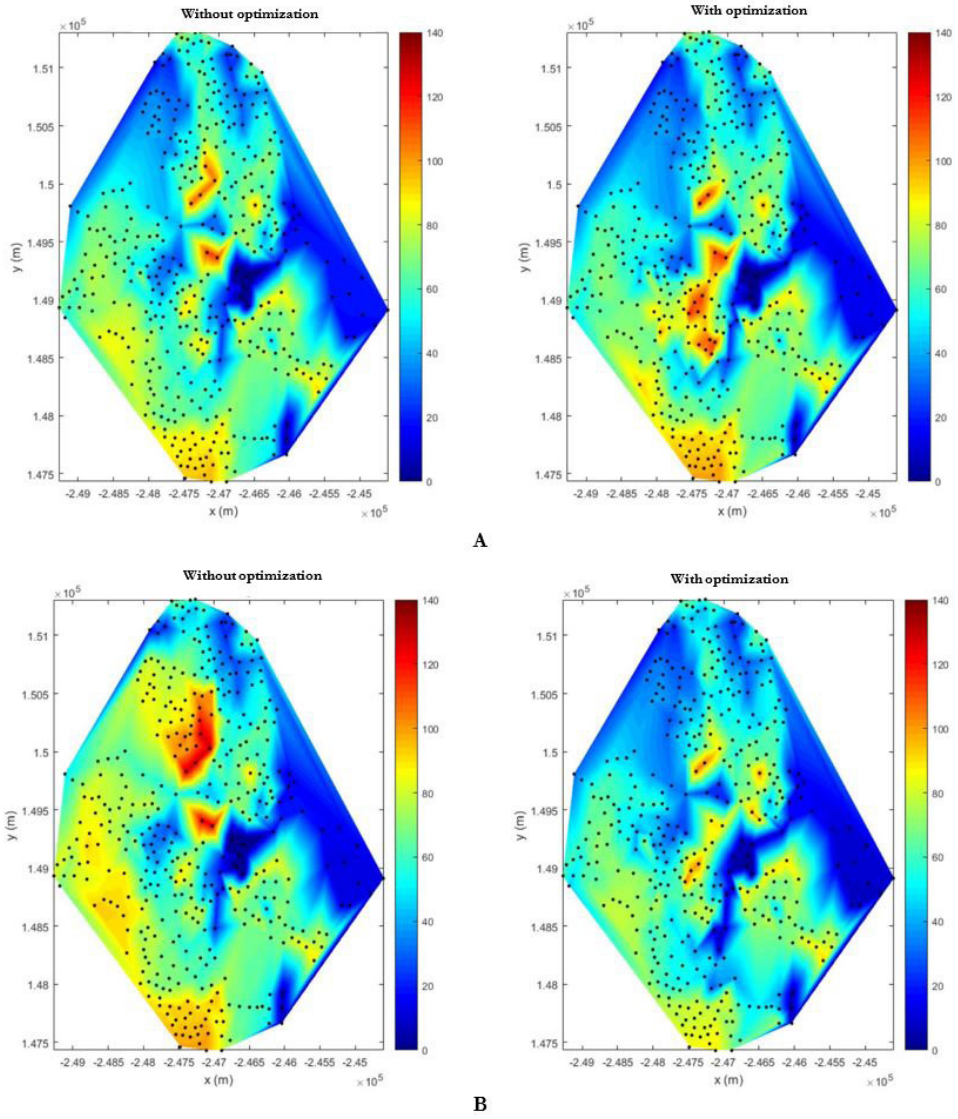

Figure 8. Distribution of the pressures in the D-Town network, without and with optimization, for the topological criterion "Length 15 km" in conditions of maximum (A) and minimum demand (B).

overall pressure required by the network distribution. From the quantitative point of view, the PU in the network was reduced from 52,07 to 44,33 in average. This reduction at the PU corresponds to a leakage reduction of 30% at the entire network. This leakage is calculated following the methodology presented by Brentan et al. (2017) and take into account a scenario of the network in operation without DMA's and the scenarios with DMAs.

Even if the benefits of DMA design are clear, knowing the diversity and dynamic of WDN, it's a hard task to evaluate how much will be this benefit for a water utility without simulations and deeper studies of particular cases.

CONCLUSION

This work presented the comparison between a hybrid model (SOM + k-means) and a k-means method model for the creation of DMAs with the purpose of optimizing the water supply system, considering the similarity of the topological conditions of the nodes of the network, mathematical criteria and topological criteria to find the optimum number of DMAs.

The topological similarity of nodes in the water distribution network was essential for the effective creation of DMAs. The k-means method performed well, presenting good quality assessment indexes and the ability to simplify the water supply network, an important feature for water distribution management.

The use of mathematical criteria by itself can generate an impractical solution from the hydraulic point of view and for future work, the topological criteria must be considered jointly with the mathematical criteria to improve the quality of the creation of DMA.

Depending on the criteria used, the size and configuration of the DMAs will be unique and it is up to the system's managers to choose the criteria that will best suit the water distribution network, considering the costs involved.

From the mathematical point of view, the DMA design process can be affected not only by the hydraulic or physical features, but for the optimization design problem. In this work, the optimization is applied for the optimal control valves placement. In this sense, the costs of the valves (related to the number of valves and diameter size) are minimized, taking into account operational parameters, such as the pressure deficit and pressure uniformity in a single-objective approach. he problem could be easily passed for a multi-objective optimization, considering the evaluation parameters (Resilience, Pressure uniformity, etc) as objectives, or becoming the constrains of the problem, in objectives to be reached. If in one hand, the multi-objective approach can be useful for real and complex problems, on the other hand, the final Pareto's front should be treated and the opinion of decision makers will play an important rule for the final solution of the problem.

REFERENCES

ALHIMIARY, H. A.; ALSUHAILY, R. H. Minimizing leakage rates in water distribution networks through optimal valves settings. In: WORLD ENVIRONMENTAL AND WATER RESOURCES CONGRESS 2007, 2007, Tampa, Florida. *Proceedings*… Reston: American Society of Civil Engineers, 2007. p. 1-13. http://dx.doi.org/10.1061/40927(243)495.

ALHONIEMI, E.; HOLLMEN, J.; SIMULA, O.; VESANTO, J. Process monitoring and modeling using the self-organizing map. *Integrated Computer-Aided Engineering*, v. 6, n. 1, p. 3-14, 1999. http://dx.doi.org/10.3233/ICA-1999-6102.

BRENTAN, B. M.; LUVIZOTTO, E.; MONTALVO, I.; IZQUIERDO, J.; PÉREZ-GARCÍA, R. Near real time pump optimization and pressure management. *Procedia Engineering*, v. 186, p. 666-675, 2017.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics*, v. 3, n. 1, p. 1-27, 1974.

CAMPBELL, E.; AYALA-CABRERA, D.; IZQUIERDO, J.; PÉREZ-GARCÍA, R.; TAVERA, M. Water supply network sectorization based on social networks community detection algorithms. *Procedia Engineering*, v. 89, p. 1208-1215, 2014. http://dx.doi.org/10.1016/j.proeng.2014.11.251.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 1, n. 2, p. 224-227, 1979. http://dx.doi.org/10.1109/TPAMI.1979.4766909. PMid:21868852.

DI NARDO, A.; DI NATALE, M. A heuristic design support methodology based on graph theory for district metering of water supply networks. *Engineering Optimization*, v. 43, n. 2, p. 193-211, 2011. http://dx.doi.org/10.1080/03052151003789858.

DI NARDO, A.; DI NATALE, M.; SANTONASTASO, G. F. A comparison between different techniques for water network sectorization. *Water Science and Technology:* Water Supply, v. 14, n. 6, p. 961-970, 2014. http://dx.doi.org/10.2166/ws.2014.046.

DIAO, K.; ZHOU, Y.; RAUCH, W. Automated creation of district metered area boundaries in water distribution systems. *Journal of Water Resources Planning and Management*, v. 139, n. 2, p. 184-190, 2012. http://dx.doi.org/10.1061/(ASCE)WR.1943-5452.0000247.

EBERHART, R. C.; KENNEDY, J. A new optimizer using particle swarm theory. In: INTERNATIONAL SYMPOSIUM ON MICRO MACHINE AND HUMAN SCIENCE, 6., 1995, Nagoya, Japan. *Proceedings*… Piscataway, New Jersey: IEEE, vol. 1, 1995. pp. 39-43. http://dx.doi.org/10.1109/MHS.1995.494215.

FAIRES, J. D.; BURDEN, R. *Numerical Methods*. 2nd ed. Pacific Grove: Brooks/Cole, 1998.

FARLEY, M. R. *District metering, part I-system design & installation*. Swindon: WRc Engineering, 1985. 84 p.

GALDIERO, E., PAOLA, F., FONTANA, N., GIUGNI, M., SAVIC, D. Decision Support System for the optimal design of District Metered Areas. *Journal of Hydroinformatics*, v. 18, n. 1, p. 49-61, 2015.

GRAYMAN, W. M.; MURRAY, R.; SAVIC, D. (2009). Effects of redesign of water systems for security and water quality factors. In: STARRETT, S. (Ed.). WORLD ENVIRONMENTAL AND WATER RESOURCES CONGRESS, May 2009, Kansas City, Missouri. *Proceedings*… Reston, VA: American Society of Civil Engineers, 2009. p. 17-21. http://dx.doi.org/10.1061/41036(342)49.

HAYKIN, S. *Redes Neurais:* princípios e prática. 2a ed. Porto Alegre: The Bookman, 2001. 902 p.

HERRERA, M.; TORGO, L.; IZQUIERDO, J.; PEREZ-GARCIA, R. Predictive models for forecasting hourly urban water demand. *Journal of Hydrology*, v. 387, n. 1-2, p. 141-150, 2010. http://dx.doi.org/10.1016/j.jhydrol.2010.04.005.

HUANG, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, v. 2, n. 3, p. 283-304, 1998. http://dx.doi.org/10.1023/A:1009769707641.

KAUFMAN, L.; ROUSEEUW, P. J. *Finding groups in data:* an introduction to Cluster Analysis. Hoboken: John Wiley & Sons, Inc., 1990. http://dx.doi.org/10.1002/9780470316801.

KOHONEN, T. *Self-Organizing Maps*. Berlin: Springer-Verlag, 2001. http://dx.doi.org/10.1007/978-3-642-56927-2.

KOUA, E. L. & KRAAK, M. J., 2006. Evaluating the usability of visualization methods in an exploratory geovisualization environment. *International Journal of Geographical Information Science*, v. 20, n. 4, p. 425-448.

MARCHI, A.; SALOMONS, E.; OSTFELD, A.; KAPELAN, Z.; SIMPSON, A. R.; ZECCHIN, A. C.; MAIER, H. R.; WU, Z. Y.; ELSAYED, S. M.; SONG, Y.; WALSKI, T.; STOKES, C.; WU, W.; DANDY, G. C.; ALVISI, S.; CREACO, E.; FRANCHINI, M.; SALDARRIAGA, J.; PÁEZ, D.; HERNÁNDEZ, D.; BOHÓRQUEZ, J.; BENT, R.; COFFRIN, C.; JUDI, D.; MCPHERSON, T.; VAN HENTENRYCK, P.; MATOS, J. P.; MONTEIRO, A. J.; MATIAS, N.; YOO, D. G.; LEE, H. M.; KIM, J. H.; IGLESIAS-REY, P. L.; MARTÍNEZ-SOLANO, F. J.; MORA-MELIÁ, D.; RIBELLES-AGUILAR, J. V.; GUIDOLIN, M.; FU, G.; REED, P.; WANG, Q.; LIU, H.; MCCLYMONT, K.; JOHNS, M.; KEEDWELL, E.; KANDIAH, V.; JASPER, M. N.; DRAKE, K.; SHAFIEE, E.; BARANDOUZI, M. A.; BERGLUND, A. D.; BRILL, D.; MAHINTHAKUMAR, G.; RANJITHAN, R.; ZECHMAN, E. M.; MORLEY, M. S.; TRICARICO, C.; DE MARINIS, G.; TOLSON, B. A.; KHEDR, A.; ASADZADEH, M. Battle of the water networks II. *Journal of Water Resources Planning and Management*, v. 140, n. 7, p. 04014009, 2013.

ROUSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, n. 1, p. 53-65, 1987. http://dx.doi.org/10.1016/0377-0427(87)90125-7.

SALDARRIAGA, J.; BOHORQUEZ, J.; CELEITA, D.; VEGA, L.; PAEZ, D.; SAVIC, D.; DANDY, G.; FILION, Y.; GRAYMAN, W.; KAPELAN, Z. Battle of the water networks district metered areas. *Journal of Water Resources Planning and Management*, v. 145, n. 4, p. 04019002, 2019. http://dx.doi.org/10.1061/(ASCE)WR.1943-5452.0001035.

SIMON, H. A. The architecture of complexity. *Proceedings of the American Philosophical Society*, v. 106, n. 6, p. 467-482, 1962.

STEINHAUS, H. Sur la division des corps matériels en parties. *Bulletin de L'academie Polonaise des Sciences*, v. 4, n. 12, p. 801-804, 1956.

SWAMEE, P. K.; SHARMA, A. K. *Design of water supply pipe networks*. New Jersey: John Wiley & Sons, Inc., 2008. http://dx.doi.org/10.1002/9780470225059.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B. Methodological*, v. 63, n. Part 2, p. 411-423, 2001. http://dx.doi.org/10.1111/1467-9868.00293.

TZATCHKOV, V. G.; ALCOCER-YAMANAKA, V. H.; BOURGUETT-ORTIZ, V. H. Graph theory based 505 algorithms for water distribution network sectorization projects. In: ANNUAL WATER DISTRIBUTION SYSTEMS ANALYSIS SYMPOSIUM WDSA, 8., 2006, Cincinnati, Ohio. *Proceeding*… Reston, VA: ASCE, 2006. p. 323-330.

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, v. 11, n. 3, p. 586-600, 2000. http://dx.doi.org/10.1109/72.846731. PMid:18249787.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; YU, P. S.; ZHOU, Z.-H.; STEINBACH, M.; HAND, D. J.; STEINBERG, D. Top 10 algorithms in data mining. *Knowledge and Information Systems*, v. 14, p. 1-37, 2008.

Authors contributions

Bernardo Novarini: main author of the paper and idea. Mainly programmer and writter.

Bruno Melo Brentan: responsible for part of codes and discussions in the article.

Gustavo Meirelles: Responsible by part of the codes and discussions in the paper.

Edevar Luvizotto Junior: supervisor and responsible by the LHC. Responsible by wirtten and revised version of the paper.