# HYBRID SOM+*k*-MEANS CLUSTERING TO IMPROVE PLANNING, OPERATION AND MANAGEMENT IN WATER DISTRIBUTION SYSTEMS

Bruno Brentan[a], Gustavo Meirelles[b], Edevar Luvizotto Junior[c], Joaquin Izquierdo[d]

[a]*brunocivil08@gmail.com*, [b]*limameirelles@gmail.com*, [c]*edevar@fec.unicamp.br*, [d]*jizquier@upv.es*

**ABSTRACT:** With the advance of new technologies and emergence of the concept of the smart city, there has been a dramatic increase in available information. Water distribution systems (WDSs) in which databases can be updated every few minutes are no exception. Suitable techniques to evaluate available information and produce optimized responses are necessary for planning, operation, and management. This can help identify critical characteristics, such as leakage patterns, pipes to be replaced, and other features. This paper presents a clustering method based on self-organizing maps coupled with k-means algorithms to achieve groups that can be easily labeled and used for WDS decision-making. Three case-studies are presented, namely a classification of Brazilian cities in terms of their water utilities; district metered area creation to improve pressure control; and transient pressure signal analysis to identify burst pipes. In the three cases, this hybrid technique produces excellent results.

**KEYWORDS:** water supply systems, classification, self-organizing maps, *k*-means clustering.

## 1. INTRODUCTION

To achieve sustainable development, cities should be able to plan, operate, and manage their infrastructure efficiently. With the growth of cities and scarcity of environmental resources, management must be re-thought to guarantee access to quality urban services for all citizens (Chouraby *et al*., 2012). Operations must be fine-tuned to this same purpose. Furthermore, planning of future government actions for urban water, such as investments to improve quality of water systems, expansion of existing systems, or reduction in energy consumption, should prioritize those cities with the worse indicators.

Many cities are unable to satisfy the needs of their populations, and smart cities respond with an intensive use of technology, information, and data to improve quality in infrastructure services. Suitably handled, this new wealth will create an ideal scenario for economic growth with a higher quality of life for citizens and less damage to the environment (Kramers *et al*., 2014). To achieve this goal, it is necessary to suitably

handle the available information about the systems, which is based on a high level of data acquisition, and thus create large databases. In water distribution systems (WDSs), these databases can integrate such variables as flow, pressure, water demand, pipe characteristics, water quality, climatic parameters, maintenance history, etc. Due to the size of the databases, computational tools are necessary to quickly identify problems and propose solutions.

Dividing the available information into clusters is a mechanism that enables identifying patterns and the main features in databases. Among several other clustering techniques, self-organizing maps (SOMs), Kohonen (1982), based on the theory of neural networks (NNs), have gained space in environmental resource research. Kohonen *et al.* (2000) highlights the use of SOMs as a clustering tool for database operation. Izquierdo *et al.* (2016) uses SOMs for early data labeling for the application of classification tools. Kalteh *et al.* (2008) highlight its extensive use in water resources problems. More specifically, in WDSs, SOMs have been widely used for water quality analysis (Blokker *et al.*, 2016), optimal design (Norouzi and Rakhshandehroo, 2011), and pattern analysis (Laspidou *et al.*, 2015), among other uses.

The application of data clustering in WDSs can help planning, operation, and management of these systems, since it is possible to identify anomalies, develop strategic operations for distribution, and evaluate the system through suitable indicators. In addition, in some cases, this classification enables the determination of interesting benchmarks which can be useful to establish quality standards and goals. We consider three situations, concisely introduced in the following three paragraphs.

Cabrera *et al.* (2014) and Lima *et al.* (2015) present different proposals for WDS classification based on the energy consumption of pumping stations to create goals for systems with poor performances. Berg and Lin (2007), Thanassoulis (2000), and Scaratti *et al.* (2013) use data envelopment analysis (DEA) to classify cities in Peru, the United Kingdom, and Brazil respectively, in relation to the management of sanitation services, and identifying places where investment is urgently needed.

Clustering applied to WDSs can also be useful to propose strategic divisions of the entire network into manageable pieces, called district metered areas (DMAs). Considering the large extension of systems and their complex interconnections, the process to divide WDSs into DMAs can be a difficult task and result in poor scenarios when partitions are developed without appropriate tools. Campbell *et al.* (2016) use social community algorithms to divide WDSs into DMAs to improve pressure control

and reduce leakage. Herrera *et al*. (2016) apply graph theory to create DMAs to improve network resilience.

Also, operational data, such as pressure or flow signals, can also be grouped to create specific strategies to suitably operate the system. Pressure signals can be used to identify burst pipes and locate anomalies in WDSs (Srirangarajan *et al.*, 2010, Covas and Ramos, 2010). Taking one of the most important challenges for WDSs, Aksela *et al.* (2012) present a methodology based on flow signal analysis with SOMs to detect leaks in WDSs.

The use of NNs and machine learning tools enables pattern extraction and classification of large databases. However, the number of resulting groups is usually too large. Classification methodologies for planning, operation, and management often require a small number of groups to ease and accelerate decision-making. SOMs are unable to create this reduced predefined number of groups, which is a drawback for their application in these cases.

When the number of groups can be predefined, the use of classical clustering methods, such as *k*-means algorithms, has been widely explored, since these methodologies can be faster than NN approaches. However, NN feature extraction combined with fast processing by *k*-means can be an interesting way to handle big data.

This paper proposes a classification method using SOMs as a pre-processing technique, coupled with the *k*-means algorithm to achieve groups that can easily be labeled and used in WDS decision-making. To exemplify the use of this methodology, three case studies are addressed. In the first, Brazilian cities are classified with respect to their water quality, operational efficiency, and economic performance, and this information can help the government plan investments to raise water quality. The second study applies the hybrid model to generate a scenario of DMAs in a fictitious water distribution network (WDN), C-town, which is frequently used as a benchmarking problem. This application enables water utilities to improve control efficiency, mainly for pressure management (which is closely related to water losses). Finally, in the third case, hydraulic transient pressure signals generated by pipe bursts and ruptures are processed with the proposed hybrid SOM+*k*-means algorithm to extract patterns and features in the data. This can help water companies to deal with anomalous events in WDSs, since knowledge of event groups can be associated with specific strategies to quickly solve the problems.

This paper is a substantial extension of Brentan *et al.* 2016, in which only a simplified version of the first case study was considered.

## 2. CLUSTERING PROCESS

### 2.1. Self-organizing maps (SOMs)

Based on brain behavior under visual and memory stimulation, SOMs are a type of neural network with unsupervised training as proposed by Kohonen (1982). When stimulated, different regions of the brain can react according to the pattern of the stimulus. This behavior enables separating different stimuli and triggering more efficient reactions. With this inspiration, an SOM is a tool to process input data and find patterns to group similar data.

SOMs have been applied in many research fields mainly because of their ability to learn from high-dimensional input data, resulting in a low-dimensional (usually bi-dimensional) output layer (Kohonen, 2000). This property helps the visualization of topological correlations among data and leads to better understanding of a problem.

A competitive learning process is responsible for the training of SOMs. Basically, this learning process is made of three steps: competition; cooperation; and synaptic update. The competition stage is responsible for identifying the map region most activated by a certain input data. Such a region can be defined by the neighborhood of the most activated neuron, the so-called winning neuron. In the cooperation stage, the influence of the winning neuron on its neighborhood is determined. Finally, in the synaptic update stage, the winning neuron increases its weight value, leading to new and similar input data to activate the neuron again (Haykin, 1990).

The iterative process of mesh adjustment to represent the feature space requires that all samples are presented to the network; in each presentation the most excited region of the mesh is adjusted. This region can be identified through the neuron exhibiting the greatest similarity between an input datum $\boldsymbol{x}$ and a neuron $\boldsymbol{w}_j$. This means the minimal distance between $\boldsymbol{x}$ and a neuron $\boldsymbol{w}_j$ selects this neuron as the most excited and initiates the process of mesh adjustment. The Euclidian distance has been widely applied to determine this similarity.

As suggested by the biological inspiration, the activation of the winner also defines a topological neighborhood that will be excited. The closer to the winning

neuron, the more excited will be its neighborhood by the input $\boldsymbol{x}$. The activation power of a neuron inside the neighborhood may be written as a monotonically decaying function, $h_{j,i(\boldsymbol{x})}$, centered on the winning neuron $i(\boldsymbol{x})$ and containing the set of $j$ neurons excited by such a winner.

The influence on the neighborhood topology reduces with time to make the model more realistic. This means that from one iteration to the next, the influence of the winning neuron decreases, making the weight adjustment more stable. After completing an iteration, each weight (neuron position) is updated with its corresponding increment $\Delta\mathbf{w}_j$ defined by (1):

$$\Delta\mathbf{w}_j = \eta h_{j,i(\boldsymbol{x})} (\mathbf{x} - \mathbf{w}_j) \tag{1}$$

where $\eta$ is the forgetfulness rate, which represents the human-like learning process.

Finally, the updating for each time step $n$ can be written as:

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n) h_{j,i(\boldsymbol{x})} (\mathbf{x} - \mathbf{w}_j(n)) \tag{2}$$

The learning process finishes when either the mesh update is smaller than a pre-defined threshold, or the number of iterations is reached. At this point, the mesh can represent the feature space in a lower dimension than that of the original space and some correlations between input data dimension can be observed. Furthermore, the input data can be classified by the nearest neuron, thus forming a set of clusters.

The quality of an SOM is linked to the capacity to represent the feature space with as little distortion as possible. This quality is closely connected to the map structure. A high number of neurons enables the data to be adjusted – however, it makes the map too rigid, harms the classification of new data, and increases computational costs. On the other hand, a map with few neurons has cheaper training process but may distort the information of the feature space. In short, map architecture definition is a not simple task.

A way to evaluate the final quality of an SOM is the quantization error that corresponds to the mean distance between the final winning neuron for each datum (Sun, 2000). This is a resolution measurement and is correlated with the number of neurons. To find an optimal configuration of the map, an index to evaluate the

efficiency of the map is presented, using the quantization error and processing time. The optimal configuration should minimize both the quantization error and the processing time. A computational index ($I_{eff,i}$) is applied to evaluate the performance of an SOM (Eq. 5). The normal processing time and normal quantization error are summed, and the minimal value of this index provides the optimal configuration.

$$I_{eff,i} = \frac{QE_i - \overline{QE}}{\sigma_{QE}} + \frac{t_i - \bar{t}}{\sigma_t}. \quad (5)$$

Here $t_i$ is the processing time for a map $i$; $\bar{t}$ is the mean time for all maps with the standard deviation $\sigma_t$; $\overline{QE}$ is the mean quantization error; and $\sigma_{QE}$ is the standard deviation for the $QE$ series. The quantization error, $QE_i$, corresponds to the mean representability of the input data $\mathbf{x}_j$ by the corresponding winning neuron $\mathbf{w}_j$, and can be written as:

$$QE_i = \frac{1}{n}\sum_{j=1}^{n}\left\|\mathbf{x_j} - \mathbf{w}_j\right\|. \quad (6)$$

However, for certain practical applications, the number of neurons obtained through SOMs can be much higher than the desirable number of clusters. To reduce the number of clusters, while taking advantage of the pre-clustering performed via SOMs, the *k*-means algorithm is applied.

### 2.2. *k*-means algorithm

Its separation ability and easy implementation makes the *k*-means methodology applicable in many research fields (Herrera *et al*., 2010, Godin *et al*., 2005, Laerhofen, 2001). In addition, as an unsupervised algorithm, the *k*-means algorithms require a previous definition of the number of groups to start the labelling data process.

Given a number of groups, *k*, each group is represented by a centroid, which is initialized randomly in the feature space. For input data with *m* features, each centroid $\mathbf{c}_k$ can be represented as:

$$\mathbf{c}_k = [c_{k1}, c_{k2} \dots c_{km}]^{\mathrm{T}}. \quad (7)$$

The data is labeled according to the distance to the centroids. The distance to each centroid is calculated for each input datum. The nearest centroid is assigned to the input datum. When all the data is labeled, the centroid position is updated. The average position of a group *k* is written in (8):

$$\mathbf{c}_k = \frac{\sum_{i=1}^{n_k} \mathbf{x}_i}{n_k}, \quad (8)$$

where $n_k$ is the number of elements belonging to group $K$.

It is expected that at the end of the process, all data is labeled and there is the maximal distance between the clusters and a minimal distance between the centroids and their corresponding data.

The simplicity of the method of grouping the data by moving the centroids along the feature space flounders since the pre-definition of the group number *k* is a hard task. Another important point is that, in contrast to the learning process with labeled data that uses statistical error measurements to evaluate the quality of process, this clustering evaluation requires specific approaches (Maulik and Bandyopadhyay, 2002).

As the definition of the number of groups is not simple, a common way to assess the quality of clustering is to apply the clustering method for different numbers of groups, and then use a quality index to evaluate the clustering performance. This quality evaluation usually considers the capacity to separate elements with maximal distance (inter-cluster criterion) and the capacity of the data to gather together around the centroids, generating maximal compact groups (intra-cluster criterion). The work by Maulik and Bandyopadhyay (2002) presents a set of validation and quality indexes. Among these indexes, an important quality indicator of clustering quality is that proposed by Calinski and Harabaz (1974) and called the $CH$ index, written as:

$$CH = \left[\frac{\sum_{k=1}^{K} n_k \|\mathbf{c}_k - \mathbf{c}\|^2}{K-1}\right] \Big/ \left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{\|\mathbf{x}_i - \mathbf{c}_k\|^2}{n-K}\right], \quad (9)$$

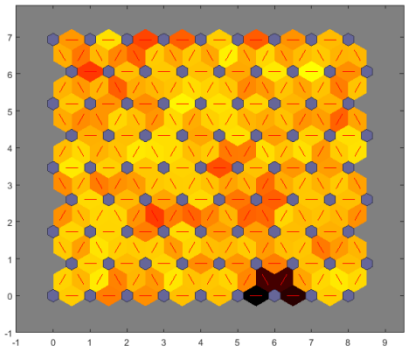
where $n_k$ is the number of elements of cluster $K$, $\mathbf{c}$ is the centroid of all input data, $K$ is the number of clusters, and *n* is the number of input data. This index takes into account

analyses of the variance method (ANOVA) and calculates centroid distances and distances between data and centroids, by correlating internal and external distances. The *CH* index can be understood as a relation between centroid distances (external evaluation) and the data clustered around the corresponding centroids (internal evaluation). This calculation allows defining the best *K* partition of data, the highest *CH* being linked to the best number of clusters.

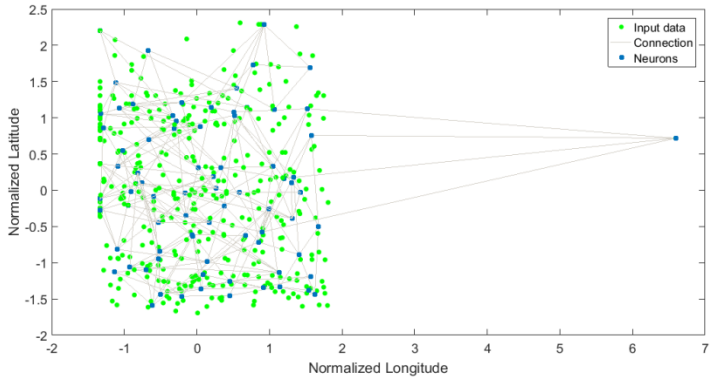
### 2.3. Hybrid methodology SOM+*k*-means

The larger the mesh of neurons used in an SOM, the larger is the number of groups resulting from the clustering process. Merging the clusters from an SOM can reduce the number of final groups. However, it can also decrease the quality of clustering if this task is done manually. Considering the application of this work, namely, to improve planning, operation, and management of WDSs, a predefined number of groups is deemed necessary.

The *k*-means algorithm is applied here to reduce the number of clusters, using the *CH* index to define the ideal number of clusters. At the end of the SOM process, all neuron positions are known and this enables a certain local representation of the data. A manner to visualize the final clustering of an SOM is the U-matrix (Siemon and Ultsch 1990) because it represents the distance between the neurons and their neighborhood. A U-matrix allows the visualization of the final configuration of the map and helps investigate the previous number of macro clusters. Figure 1a illustrates the U-matrix for the DMA creation data, a case study of this work. Light colors represent small distances between neurons, and the darker the color, the larger the distance between neurons. Even if an isolated neuron is highlighted, a clear cluster region is not evident. The final neuron positions in the output space (Figure 1b) can be used to re-cluster the data to explore features of the input space and create well defined clusters.




**a)** U-matrix for DMA creation data after the optimal architecture definition




**b)** Final neuron position after training for the optimal SOM.

**Figure 1**. Final results after training an SOM that highlight the neuron positions and data

Identification of the number of groups is not easy. However, considering the data and neurons in the feature space, it is possible to calculate the distance between neurons and data. This distance indicates how large is the similarity between a datum and a neuron. Small values for the distance between a neuron and an input datum indicate a good representativeness of the data by the neuron.

The distance between neurons and data can be calculated and organized in a matrix (a dissimilarity matrix) which is a manner to interpret SOM results, since all distances between neurons and data are known. Furthermore, by using the dissimilarity matrix as input data for *k*-means, the predefined groups obtained from the SOM will be considered, resulting in a good option to merge SOM groups, and eventually reduce the final number of groups. The dissimilarity matrix may be written as:

$$D = \begin{bmatrix} \|w_1 - x_1\|^2 & \dots & \|w_1 - x_i\|^2 \\ \vdots & \ddots & \vdots \\ \|w_l - x_1\|^2 & \dots & \|w_l - x_i\|^2 \end{bmatrix}. \quad (10)$$

Once the dissimilarity matrix is calculated, it is possible to apply *k*-means clustering for a range of clusters and evaluate the final results via the *CH* index. The best value for the *CH* index will define the number of groups and thus, the clustering scenario can be evaluated in terms of the features of each group. The flowchart in Figure 2 shows the entire hybrid process.

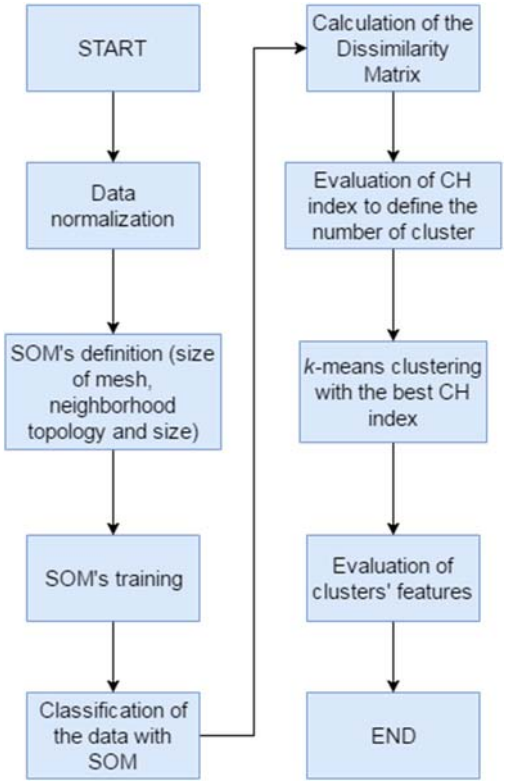



**Figure 2**. Entire clustering process using the hybrid SOM+*k*-means model

## 3. CASE STUDY 1: INDICATORS FOR WATER SUPPLY PLANNING

### 3.1. Case study description

The Brazilian National System of Sanitation Information (SNIS) is a large database on sanitation service performance. It is an important tool for planning public policies and management of the investments made. However, the large volume of information makes an assessment of the necessary improvements difficult.

To improve the global quality of water services, the investment fields were divided and three major groups of indicators (namely: water quality, operational efficiency and economic performance) were created, each containing a set of relevant indicators for the global system quality evaluation described in Table 1.

Although the fulfillment of SNIS is mandatory for water utilities in Brazil, some cities omitted information or presented data with errors, compromising the evaluation of all Brazilian cities. To obtain consistent results, the cities with these data problems were disregarded. The most recent available data, used in this work, is from 2014. Cities with missing information or wrong values for the indicators were disregarded at a pre-processing stage. Most of the indicators used in this work are expressed in percentage values. As a result, 2231 cities are considered for the study, which represents 40% of

the total number of cities in Brazil. As many cities are disregarded due to missing information, it is possible that critical cities are not considered in this work. After applying the methodology proposed here, the data is normalized using the *zscore* function, which can set the data in a small range, while maintaining the distribution.

**Table 1.** Indicators selected for water supply classification

| | Indicator | Description |
|---|---|---|
| **WATER QUALITY** | - | Does a wastewater service exist? |
| | IN055_AE | Water service index (%) |
| | IN057_AE | Water fluoridation index (%) |
| | IN079_AE | Samples conformity - residual chlorine (%) |
| | IN080_AE | Samples conformity - turbidity (%) |
| | IN085_AE | Samples conformity - total coliforms (%) |
| **OPERATIONAL EFFICIENCY** | IN009_AE | Micro metering index (%) |
| | IN011_AE | Macro metering index (%) |
| | IN049_AE | Leakage index (%) |
| | IN055_AE | Water service index (%) |
| | QD003 | Duration of stoppages (h/year) |
| | QD022 | Duration of interruptions (h/year) |
| **ECONOMIC PERFORMANCE** | - | Relation between investment and revenue (%) |
| | IN005_AE | Water tariff ($/m$^3$) |
| | IN019_AE | Relation between active economies and employees (econ/empl) |
| | IN026_AE | Exploitation expenses ($/m$^3$) |
| | IN036_AE | Expenses with employees (%) |
| | IN037_AE | Expenses with electrical energy (%) |
| | IN038_AE | Expenses with chemical products (%) |

### 3.2. Results and discussion

The first clustering level, performed with an SOM, has the architecture defined by an optimal analysis of the computational index presented in equation (5). The variation in the number of neurons decreases significantly the quantization error because the larger the mesh, the smaller is the distance between the winning neuron and the datum. The size of the mesh is closely linked to the processing time, which increases with the number of neurons.

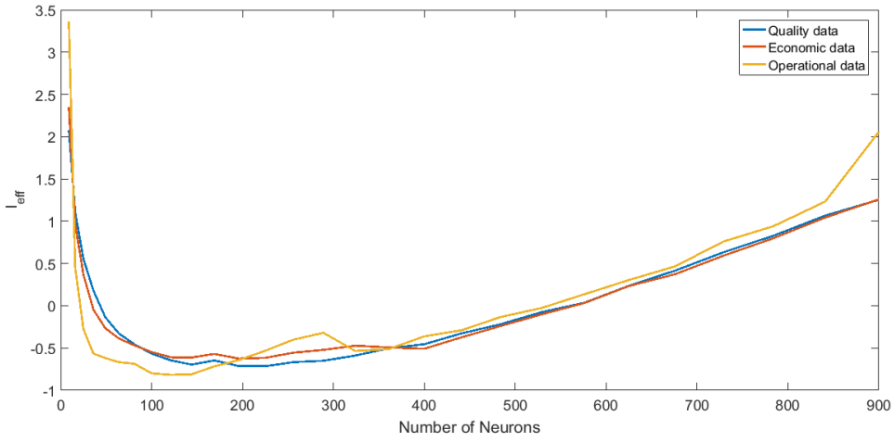
The SOM is computed by varying the number of neurons from 9 to 900 and the size of the neighborhood from 1 to 5 neurons. The architectural test and performance evaluation were developed in Matlab. Figure 3a, presents the computational index for the three indicator groups, highlighting the optimal number of neurons. Taking these

values for water quality: SOMs with 144 neurons (operational data), 169 neurons (economic data), and 81 neurons were implemented. The definition of neighborhood shows little importance in the final results in terms of improvement of quantization error.

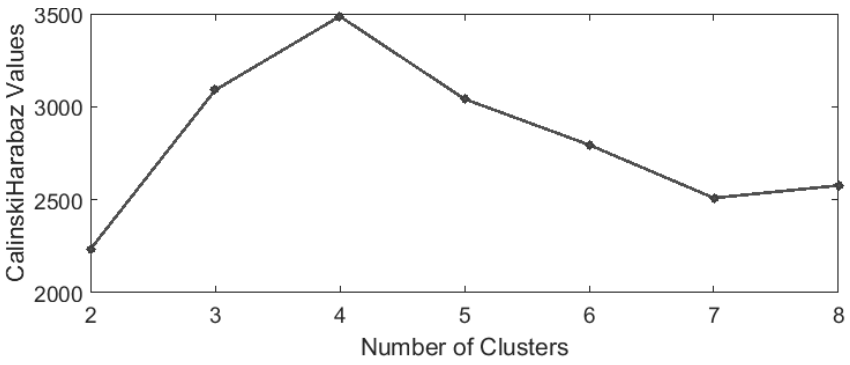
The final results of the SOMs are computed in the dissimilarity matrix. In this case, taking the 81 neurons and the 2231data, the dissimilarity matrix can be written as:

$$D = \begin{bmatrix} \|w_1 - x_1\|^2 & \dots & \|w_1 - x_{2231}\|^2 \\ \vdots & \ddots & \vdots \\ \|w_{81} - x_1\|^2 & \dots & \|w_{81} - x_{2231}\|^2 \end{bmatrix}. \quad (11)$$

This dissimilarity matrix $D$ is the input for the $k$-means process. The number of final clusters is defined by the best value of the *CH* index for data quality – as shown in Figure 3b. Operational and economic data follows the same trends of quality analysis and also results in four groups. The evaluation of the best final number of clusters was made in the range of 2 to 7 groups, as indicated in Figure 3b. This *a priori* limitation of the number of groups avoided the generation of too many groups, which could spoil decision-making effectiveness about the management of the cities.



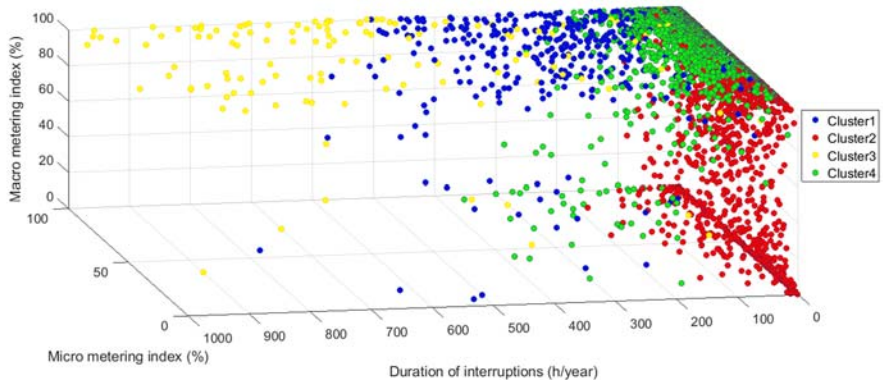

**a)** $I_{eff}$ to define the optimal architecture of SOM



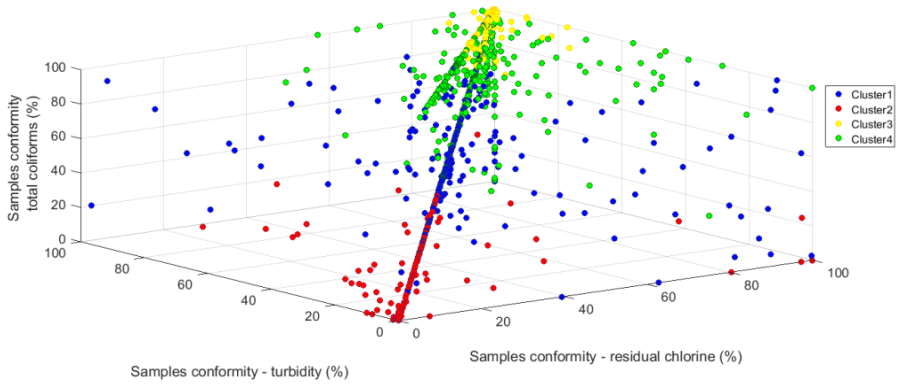

**b)** CH index for water quality data

**Figure 3**. Computational evaluation of SOM+*k*-means algorithm to define the SOM architecture (a), and the number of groups for *k*-means (b)

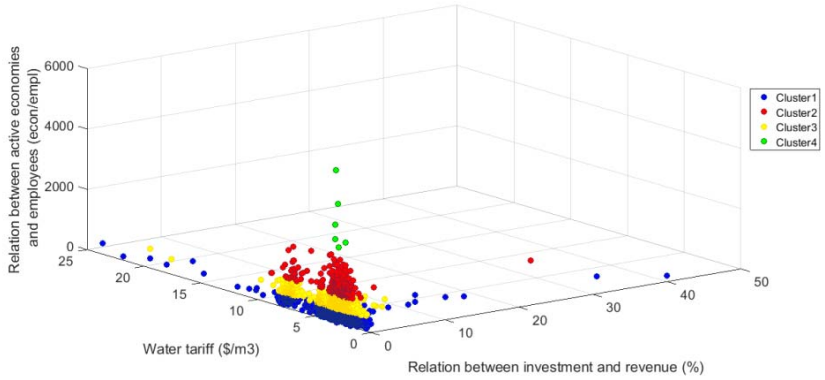
As the three groups have more than three dimensions, 3D representations of clusters are not completely accurate. However, it is possible to observe the main features of the clusters. The most representative figures for the three-dimensional plotting of clustered data are presented in Figure 4. Operational indicators (Figure 4-a) have the most mixed representation, while economic performance indicators (Figure 4-c) have a very well defined scattering.




**a)** Operational efficiency clusters




**b)** Water quality clusters




**c)** Economic performance clusters

**Figure 4.** Final clustering for water supply performance considering the three main groups of indicators

By calculating the average for each group (Table 2), some noticeable features can be observed: for water quality indicators, the best groups contain the majority of cities (which can be explained by the relation with health issues and recent investments to improve sewage collection and treatment aimed at reducing waterborne diseases) (Gonçalves, 2014). The other indicators follow a strict correlation with wastewater service availability. The larger the water service, the larger the mean of cities with wastewater services and the better the quality indicators. Furthermore, the fluoridation index also has an important correlation with the other quality indicators, such as the total number of coliforms. The worst cluster, Cluster 2, presents the lowest value for water services and includes most of cities without wastewater services, leading to alarming values for quality indicators.

**Table 2.** Average values for each group

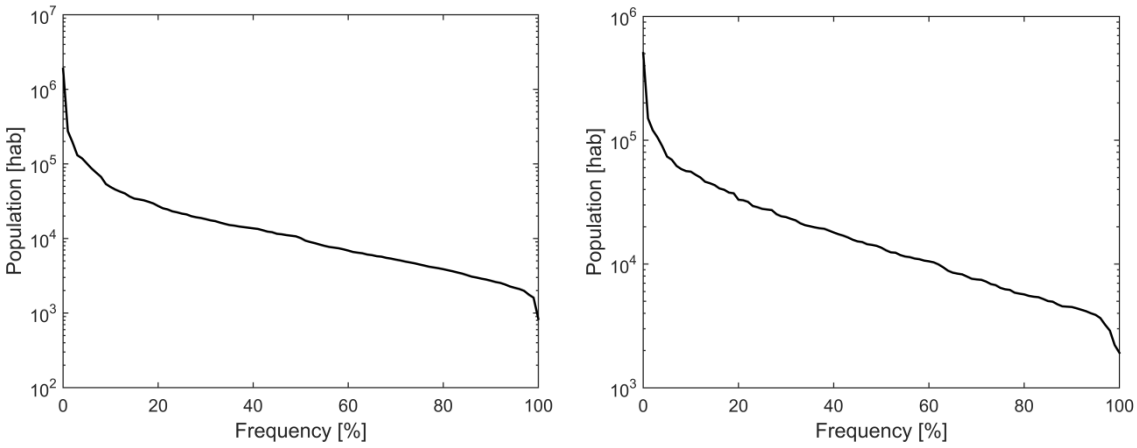
| | Group | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **WATER QUALITY** | - | 0.27 | 0.11 | *0.60* | 0.38 |
| | IN055_AE | 58.2 | 41.0 | *80.8* | 68.6 |
| | IN057_AE | 10.4 | 0.0 | *97.1* | 26.7 |
| | IN079_AE | 54.4 | 10.9 | *98.4* | 88.0 |
| | IN080_AE | 50.9 | 10.0 | *98.1* | 86.4 |
| | IN085_AE | 54.4 | 10.7 | *99.0* | 92.5 |
| | Population | 23,937 | 25,134 | *26,619* | 23,940 |
| | Revenue [$] | 2,313,207 | 914,278 | *6,632,448* | 3,047,526 |
| | Number of cities [%] | 15.22 | 13.45 | *37.31* | 34.03 |
| **OPERATIONAL EFFICIENCY** | **Group** | **1** | **2** | **3** | **4** |
| | IN009_AE | *94.1* | 90.0 | 89.9 | 84.7 |
| | IN011_AE | *72.9* | 62.7 | 62.7 | 57.1 |
| | IN049_AE | *34.8* | 30.9 | 29.8 | 30.3 |
| | IN055_AE | *72.3* | 71.2 | 70.1 | 68.1 |
| | QD003 | *78* | 254 | 102 | 195 |
| | QD022 | *45* | 617 | 243 | 665 |
| | Population | *35,048* | 40,840 | 35,092 | 42,215 |
| | Revenue [$] | *6,293,141* | 10,175,034 | 7,444,753 | 10,291,823 |
| | Number of cities [%] | *3.32* | 21.44 | 21.80 | 53.44 |
| **ECONOMIC PERFORMANCE** | **Group** | **1** | **2** | **3** | **4** |
| | - | 0.16 | *0.18* | 0.11 | 0.02 |
| | IN005_AE | 3.3 | *3.2* | 3.4 | 2.9 |
| | IN019_AE | 285 | *1,208* | 636 | 3,538 |
| | IN026_AE | 3.7 | *2.3* | 2.9 | 1.3 |
| | IN036_AE | 66.5 | *60.4* | 64.5 | 46.3 |
| | IN037_AE | 16.8 | *22.0* | 18.9 | 30.6 |
| | IN038_AE | 2.6 | *2.1* | 2.5 | 2.5 |
| | Population | 34,042 | *93,172* | 51,846 | 26,940 |
| | Revenue [$] | 6,784,361 | *24,699,491* | 14,320,835 | 3,753,564 |
| | Number of cities [%] | 59.61 | *8.02* | 32.23 | 0.14 |

Regarding the operational efficiency group, the best group found contains a minority of the cities. In both cases, it is observed that the best groups have higher revenues, but the size of the cities does not appear to influence performance. A correlation with the population (and consequently the size of the WDS) and stoppages and interruptions is also noted. The best equipped group in terms of water metering (micro and macro) has the worst leakage index. This correlation, which sounds contradictory, can be linked to a better knowledge about leakage by water utilities, while the other groups, with lower metering indexes, are only estimating their leakage index, thus generating inaccurate values.

Considering economic performance, it is clear that cities that have higher investments rates also have higher revenues and population. The indicators show that for this group, the amount collected is used efficiently in general expenses as

employees, energy, and chemical products. However, these cities seem to be benefit from a favorable topology of their networks, as exploitation expenses are low.

When best and worse groups are evaluated, it is observed that they contain a good distribution of cities of every size (Figure 5). Therefore, it is plausible to conclude that good planning and management depends mainly on short, medium, and long-term strategic actions, and efficiently using the available resources on priority areas, which can be defined by the presented clustering.




**(a)** **(b)**

**Figure 5.** Frequency of cities size: a) best groups; b) worst groups.

## 4. CASE STUDY 2: DMA CREATION CRITERIA FOR PRESSURE MANAGEMENT

### 4.1. Case study description

WDSs are usually compounded by a large set of pipes and devices. The larger the network, the more complex is its management, since the topology and topography can change significantly, turning pressure management into a hard task.

Several authors have proposed methodologies to segregate WDNs into smaller pieces, called DMAs (Campbell *et al*., 2016; Di Nardo *et al*., 2011). The division of the entire network aims at the creation of DMAs with similar features, such as elevation and demand, to ease management by water utilities. Furthermore, DMA creation can help water utilities in an important point of management: leakage reduction. When a DMA is created, flowmeters and pressure reducing valves (PRVs) can be installed at the

entrance to monitor the leakage level in the DMA and accurately control pressure (Araujo *et al.*, 2006).

After the grouping of nodes, the management of DMAs requires the installation of isolation valves and the closure of the pipes that interconnect different DMAs. The closure of pipes creates preferred ways to deliver the water from sources to sectors. The choice of entrances for the DMA can be looked on as an optimization problem, since PRV installation depends on the cost and operational conditions of the system (Galdiero *et al*., 2016). Brentan *et al*. (2017) present a multi-level optimization methodology to identify the optimal entrance of the DMAs and the optimal set-point for valves. The authors apply particle swarm optimization (PSO) combined with hydraulic simulations to reach the minimal cost for the controlling device installation. This methodology is applied in this work to the final result produced by the hybrid SOM+*k*-means clustering proposed in this paper.

### 4.2. Results and discussion

To this purpose, topological features (nodal demand and elevation) and the spatial position of the node are used as input data for SOM. The study is applied on the C-Town network, a benchmarking case study used in many WDS analyses, such as (Marchi *et al*., 2012, Brentan *et al*., 2017). The C-Town has 398 nodes linked by 458 pipes. Users are supplied from a reservoir and seven tanks that use 13 pumps to distribute the water. The operational conditions from Wu *et al.* (2012) were considered for this study. To apply the hybrid SOM+*k*-means method, a database with the topology features of the nodes was created. The position of the node in the space, corresponding to the pair (*x*,*y*) coordinates, nodal demand, and elevation, formed the input matrix for SOM. In this case, the input matrix was also normalized by the *zscore* function, but the position data (*x*,*y*) was weighted after the normalization to increase the significance of this data in the clustering. This is important because for DMA creation, the neighborhood relationship of the nodes is paramount.

Following the procedure to find the optimal SOM architecture using the minimal computational index, an NN with 81 neurons was produced. The results of the clustering process are presented in Figure 5, considering the final number of groups by the *CH* index evaluation, which varied from 280.27 for the worst case (two groups) to 417.59 for the best case (four groups).

It is possible to observe the generation of four DMAs, maintaining the neighborhood of nodes (Figure 6). However, the scenario resulting from the hybrid method is not the most common approach, since the striped scenario requires more investment to generate isolation areas. In this case, it is possible to identify 26 boundary pipes. Table 3 presents the main features (elevation and demand) for each DMA and the number of nodes per DMA. Even if the number of nodes of each DMA is not similar, it is possible to observe the difference among the four elevations.

The number of nodes per DMA and demand do not follow a linear correlation, pointing to a weighted segregation of the network in terms of demand and elevation. Demand appears with two main groups (DMA1 and DMA2 in a group, and DMA3 and DMA4 in another group). Considering pressure management, it can be observed that elevation was an important parameter, which is good since DMAs with nodes with similar elevations are easily controlled with a pressure regulation at the entrance.

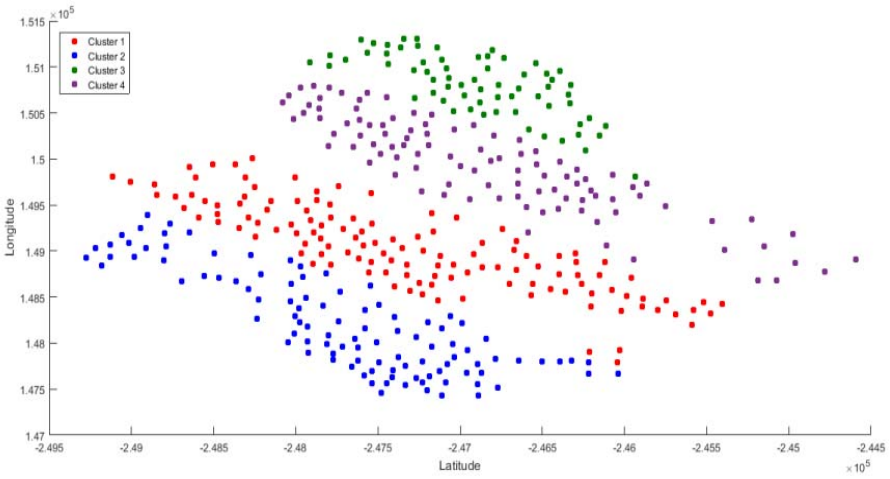



**Figure 6.** Final scenario of DMA creation using hybrid method SOM+*k*-means

**Table 3.** Average features for each DMA created using the hybrid methodology

| | DMA1 | DMA2 | DMA3 | DMA4 |
|---|---|---|---|---|
| **Mean elevation (m)** | 40.54 | 20.11 | 75.27 | 63.10 |
| **Stand. dev. of elevation (m)** | 17.53 | 13.06 | 17.94 | 14.74 |
| **Total demand (l/s)** | 131.59 | 125.38 | 88.87 | 76.43 |
| **Stand. dev. of demand (l/s)** | 0.68 | 0.71 | 0.87 | 0.69 |
| **Number of nodes** | 127.00 | 104.00 | 66.00 | 101.00 |

The pipes that separate DMAs, called boundary pipes, must be identified to allow the full isolation of DMAs. These boundaries should be manageable in order to

close or open when required. Moreover, to maximize the benefits of WDS segregation, all the entrances of a DMA should be monitored and controlled (equipping all boundary pipes with monitoring and control devices is expensive).

The result of the first optimization stage defined nine pipes with PRVs installed and 17 closed pipes (with an investment of $55,732). After defining the entrance of the DMAs, the control setting of the PRVs should be defined. In this case, the PSO is applied again while considering the nine PRVs with the aim of reducing the operational pressure of the system and so achieving the minimal required pressure (as observed in Table 4 where the pressure indicators are presented). Moreover, these indicators enable defining three main pressure zones: one formed by DMA1 and DMA2 (where the pressure uniformity parameters are similar), and the other two zones integrated by DMA3 and DMA4 (which have different values for pressure uniformity). The last column of Table 4 presents the mean values for the pressure indicator for the original C-Town network. A slight reduction of pressure uniformity and mean pressure in the network is observed. This reduction during the operation of the network can cause a significant reduction in leakage and the system thus becomes more sustainable.

**Table 4.** Pressure indicators to evaluate the DMA scenario obtained from SOM+*k*-means

| | **DMA1** | **DMA2** | **DMA3** | **DMA4** | **Mean** | **Mean (Wu *et al*, 2012)** |
|---|---|---|---|---|---|---|
| **Pressure uniformity** | 1026.12 | 1161.61 | 914.16 | 627.42 | 932.32 | 972.85 |
| **Mean pressure (m)** | 58.83 | 61.10 | 50.94 | 45.02 | 53.97 | 55.52 |
| **Minimal pressure (m)** | 25.05 | 25.08 | 25.54 | 25.33 | 25.25 | 26.69 |

## 5. CASE STUDY 3: PRESSURE SIGNAL OF HYDRAULIC TRANSIENT FLOW BY PIPE BURST AND RUPTURE

### 5.1. Case study description

Pipe bursts cause a pressure surge in the entire network. Monitoring the pressure signal in some points can help to quickly identify the area where the rupture occurred. This can be achieved with the classification of these signals, which will be different in two aspects: *pressure amplitude*, which is related to the intensity of the leakage flow resulting from the rupture; and, *time delay* for the sensor to receive this signal, which is related to the distance between the sensor and the rupture location. It is expected that

each group in the classification has a characteristic pressure signal, which could be translated to the area and intensity of the rupture, helping ensure (for example) a rapid maintenance repair and protective operations for valves and pumps.

The method of characteristics (MOC) was used to model this burst pipe scenario in WDNs. The continuity law is applied to a generic node and MOC positive and negative lines are used to calculate flow in convergent and divergent pipes respectively (Almeida and Koelle, 1992). To model a pipe burst, a sudden leakage flow is added to the nodal demand. To create a more realistic scenario, the model proposed by Van Zyl (2014) was used to simulate the leakage flow.

The hybrid method for clustering was applied to a real Brazilian DMA with 118 nodes and 153 pipes. This DMA, known as Campos do Conde II, is a part of a new system recently finished in the WDS of Piracicaba, in the State of São Paulo, Brazil. Figure 7 presents the network topology and the elevation of the nodes. Red circles highlight the monitored nodes (chosen following a pressure sensitivity analysis).

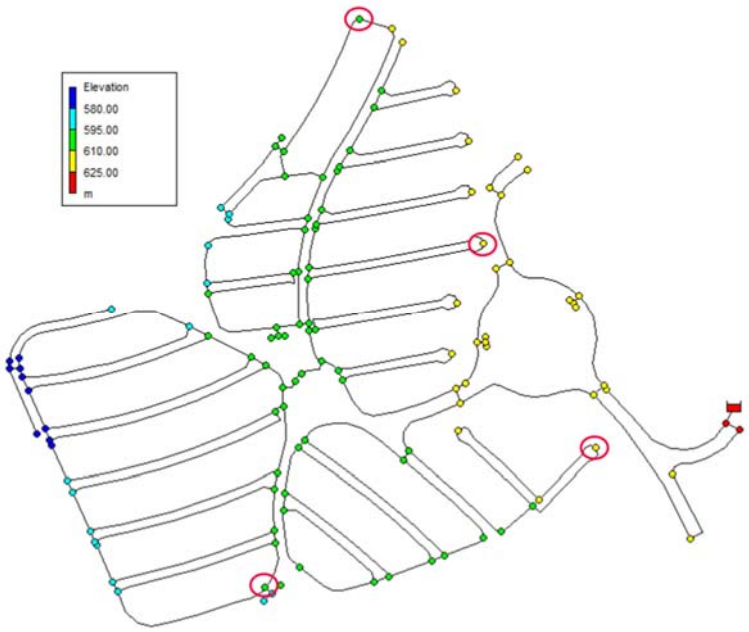



**Figure 7.** Topology of the real DMA network Campos do Conde, Piracicaba, Brazil

To create the database, simulations were performed, adding a leakage flow to each node one at a time. To create different leakage flows, the discharge coefficient was modified. Therefore, with 15 leakage conditions in each node, a total of 1,770 simulations were made.

The main objective of applying the hybrid SOM+*k*-means algorithm is to identify patterns within the set of monitored pressures in previously defined nodes. For each leakage scenario, the pressure signal of the four monitored nodes was collected during a period of 60 s and concatenated for classification with the proposed method. For this application, the data is normalized using the *z-score* function, which, in this case, has a salient feature because maintaining the signal shape is very important for a good segregation of transient pressure.

### 5.2. Results and discussion

The results of this case study were obtained with an SOM composed of 64 neurons, corresponding to the minimal value of $I_{eff}$. The application of the *CH* index analysis in the *k*-means stage provides the optimal number of clusters, ten in our case, as shown in Figure 8.

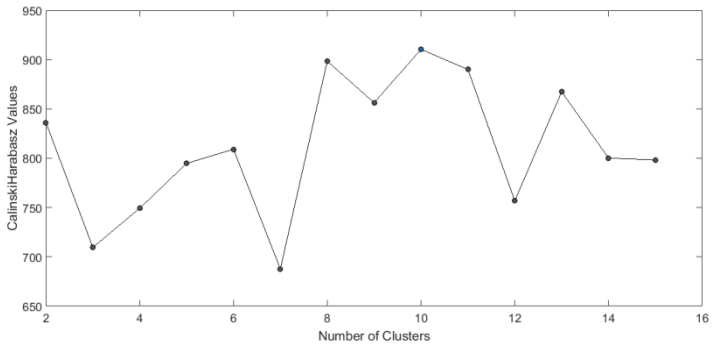



**Figure 8.** CH index for transient pressure analysis

The ten groups of pressure signals are presented and described in Figure 9. Note that each cluster contains a set of leakage scenarios with a similar pressure signal shape, which is helpful to locate the event. Clusters 2, 4, 6 and 8 also have larger pressure surges, indicating potentially harmful events. Cluster 1 is the group with most events, but mainly with minor leakages and distant from monitored nodes; while cluster 4 refers only to leakages in node 79, located near one of the monitoring nodes (creating a unique shape for the pressure signal).

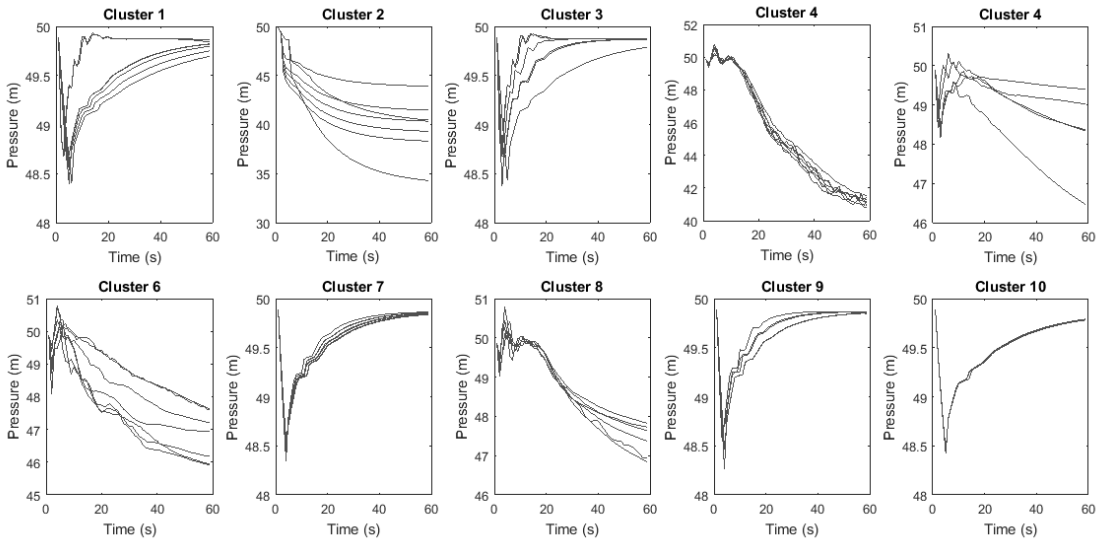



**Figure 9.** Characteristic pressure signal for each final classification cluster

The comparison presented in Figure 10 shows that for each monitoring node, the pressure signal is characteristic for each cluster. Therefore, it is expected that in near-real time operation, when a certain response is observed, according to the cluster where it is classified, predefined optimized actions may be taken to quickly solve the problem and reduce economic, environmental, and social losses.

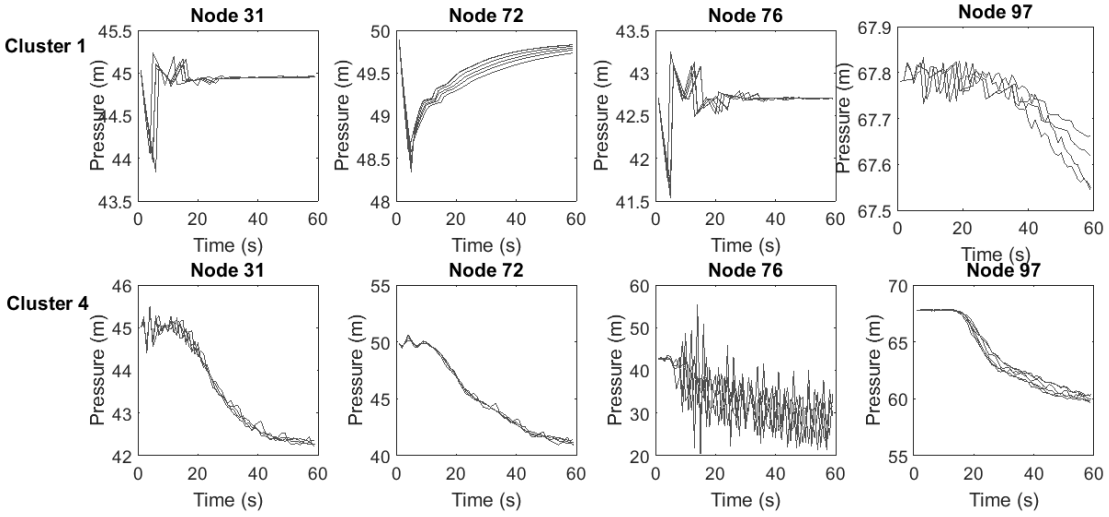



**Figure 10.** Comparison of pressure signal in monitored nodes between clusters 1 and 4

## 6. CONCLUSIONS

This paper presents a clustering tool based on SOMs coupled with a *k*-means algorithm to achieve a predefined number of groups, taking into account the complex relationship among features. In general, clustering data is a hard task, since the size of the databases and unclear relationships among their variables can hide potential clusters, thus forcing the application of robust computational tools. SOMs find patterns and project high dimensional databases in (mainly) two-dimensional maps. However, for certain applications, those maps exhibit level of granularity that is too fine. In other words, they produce a detailed description that is not directly useful in some decision-making processes and so some other procedure may be required. In this paper, we propose the hybrid use of SOMs as a preprocessing technique. Firstly, an SOM produces a feature map with the optimal architecture reached by the minimization of the efficiency index. Working on an SOM-based dissimilarity matrix, a *k*-means algorithm then coarsens the map to produce a smaller number of groups that is more suitable for various decision-making processes. Three case studies for different problems of WDSs are presented.

For the WDS indicator case study, some cities had missing or incorrect information (which could generate inconsistencies in clustering and so were discarded), showing the importance of pre-processing to obtain a reliable database. With the database of clean data, the method clearly separated the best and worst cities into different groups, which is very helpful for governmental investment policies. In addition, it is interesting to observe that the size of the city does not appear to influence WDS quality.

The method also presented good results in the DMA creation case by creating groups with similar elevation, which is a major characteristic for pressure management. The process of finding optimal entrance pressures demonstrates the importance of DMA creation due to pressure regularization. After this process, it is observed that the groups created by the hybrid method can also be differentiated by the pressure uniformity indicator, which can be used in leakage problems (since leaks are directly related to the operational pressure).

Finally, the third case study classified transient pressure signals generated by pipe burst simulations. Once again, the method presented good results, creating clusters of pressure signals with similar characteristics. It is possible to identify a class of rupture and estimate its location by just monitoring the pressure in some strategic

points, thus enabling water utilities to develop specific alarm and emergency protocols to minimize the effects of pipe bursts.

As shown in this work, management, operation, and planning of WDSs can be improved if the available data is suitably used. Moreover, the larger the database, the harder the task of applying this information from a practical point of view. The good results obtained with these three case studies (referring to planning, operation, and management of WDSs) suggests that the hybrid method proposed can also be useful to handle problems in other contexts – especially in environmental fields.

## 7. ACKNOWLEDGEMENTS

This work is partially supported by Capes and CNPq, Brazilian research agencies. The use of English was revised by John Rawlins.

## 8. REFERENCES

1. Aksela, K., Aksela, M., Vahala, R., 2009. Leakage detection in a real distribution network using a SOM. Urban Water J. 6(4), 279-289.
2. Almeida, A., Koelle, E., 1992. Fluid transients in pipe networks. Comput. Mech., Southampton, London.
3. Araujo, L., Ramos, H., Coelho, S., 2006. Pressure control for leakage minimisation in water distribution systems management. Water Resour. Manag. 20(1), 133-149.
4. Berg, S., Lin, C., 2007. Consistency in Performance Rankings: The Peru Water Sector, J. Appl. Econ. 40(6), 93-805.
5. Blokker, E., Furnass, W., Machell, J., Mounce, S., Schaap, P., Boxall, J., 2016. Relating Water Quality and Age in Drinking Water Distribution Systems Using Self-Organising Maps, Environments 3(10), 1-17.
6. Brentan, B., Meirelles, G., Luvizotto Jr., Izquierdo, J., Pérez-García, R., 2016. Water supply systems classification for wáter quality improvement, In: Sauvage, S., Sáchez-Pérez, J., Rizzoli, A., International Environmental Modelling and Software Society, V.3, pp 635-642.
7. Brentan B., Campbell, E., Meirelles, G., Luvizottor Jr., E., Izquierdo, J., 2017. Social network community detection for DMA creation: criteria analysis through multilevel optimization, Math. Probl. Eng. Article ID 9053238.

8. Cabrera, E., Cabrera Jr, E., Cobacho, R., Soriano, J., 2013. Towards an energy labelling of pressurized water networks. In: 12th International Conference on Computing and Control for the Water Industry - CCWI, Perugia - Italy.
9. Calinsky, T., Harabasz, J., 1974. A dendrite method for cluster analysis, Commun. Stat. Theory 3(1), 1-27.
10. Campbell, E., Izquierdo, J., Montalvo, I., Pérez-García, R., 2016. A Novel Water Supply Network Sectorization Methodology Based on a Complete Economic Analysis, Including Uncertainties. Water 8(5), 179.
11. Campbell, E., Izquierdo, J., Montalvo, I., Ilaya-Ayza, A., Pérez-García, R., Tavera, M., 2016. A Flexible Methodology to Sectorize Water Supply Networks Based on Social Network Theory Concepts and on Multi-objective Optimization. J. Hydroinform. 18(1), 62-76.
12. Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J., Mellouli, S., Nahon, K., Scholl, H., 2012. Understanding smart cities: An integrative framework. In Proc. 45th Hawaii International Conference on System Sciences, pp. 2289-2297.
13. Covas, D., Ramos, H., 2010. Case studies of leak detection and location in water pipe systems by inverse transient analysis. J. Water Resour. Pl. Manag. 136(2), 248-257.
14. Di Nardo, A., Di Natale, M., 2011. A heuristic design support methodology based on graph theory for district metering of water supply networks. Eng. Optimiz. 43(2), 193-211.
15. Galdiero, E., De Paola, F., Fontana, N., Giugni, M., Savic, D. 2016. Decision support system for the optimal design of district metered areas. J. Hydroinform. 18(1), 49-61.
16. Godin, N., Huguet, S., Gaertner, R., 2005. Integration of the Kohonen's self-organizing map and k-means algorithms for the segmentation of the AE data collected during tensile tests on cross ply composites, NDT & E Int. 38(4), 299-309.
17. Gonçalves, S., 2014. The effects of participatory budgeting on municipal expenditures and infant mortality in Brazil. World Dev. 53, 94-110.
18. Herrera, M., Canu, S., Karatzoglou, A., Pérez-García, R., Izquierdo J., 2010. An approach to water supply clusters by semi supervised learning, In Proc. International Environmental Modelling and Software Modelling for Environment's Sake. Otawa, Canada.

19. Izquierdo, J., Campbell, E., Montalvo, I., Pérez-García, R., 2016. Injecting problem-dependent knowledge to improve evolutionary optimization search ability. J. Comput. Appl. Math. 291, 281-292.
20. Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environ. Modell. Softw. 23, 835-845
21. Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Bio. Cybern. 43(1), 59-69.
22. Kohonen, T., 2000. Self Organization of a Massive Document Collection. IEEE T. Neural Networ. 11(3).
23. Kramers, A., Höjer, M., Lövehagen, N., Wangel, J., 2014. Smart sustainable cities–Exploring ICT solutions for reduced energy use in cities. Environ. Modell. Softw. 56, 52-62.
24. Laspidou, C. Papageorgiou, E. Kokkinos. K. Sahu, S., Gupta A., Tassiulas, L., 2015. Exploring patterns in water consumption by clustering, Procedia Engineer. 119, 1439 – 1446.
25. Lima, G., Viana, A., Dias Jr., R., Luvizotto Jr, E., 2015. Classification of water supply systems based on energy efficiency, Water Sci. Technol. 15 (6), 1193-1199.
26. Marchi, A, et al., 2012 Battle of the water networks II. Journal of water resources planning and management 140.7
27. Maulik, U., Sanghamitra B., 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE T. Pattern Anal. 24(12), 1650-1654.
28. Norouzi, K., Rakhshandehroo, G.R., 2011. A self-organizing map based hybrid multi-objective optimization of water distribution networks, T. Civil Environ. Eng. 35(C1), 105-119.
29. Siemon, H., Ultsch, A., 1990. Kohonen networks on transputers: Implementation and animation. In International Neural Network Conference (pp. 643-646). Springer Netherlands.
30. Scaratti, D., Michelon, W., Scaratti, G., 2013. Evaluation of management efficiency of the municipal services of water supply and sewage using Data Envelopment Analysis (in Portuguese). Eng. Sanit. Ambient. 18(4) 333-340.
31. Srirangarajan, S., Iqbal, M., Lim, H., Allen, M., Preis, A., Whittle, A., 2010. Water main burst event detection and localization. In Proc. Water Distribution Systems Analysis 2010, Tucson, Arizona, USA, pp. 1324-1335.

32. Sun, Y., 2000. On quantization error of self-organizing map network. Neurocomputing 34(1), 169-193.
33. Thanassoulis, E., 2000. The use of data envelopment analysis in the regulation of UK water utilities: water distribution. Eur. J. Oper. Res. 126(2), 436-453.
34. Van Zyl, J., 2014. Theoretical Modeling of Pressure and Leakage in Water Distribution Systems, Procedia Engineer. 89, 273-277.
35. Wu, Z., Elsayed, S., Song, Y. 2012. High performance evolutionary optimization for Battle of the Water Network II. Proc., 14th Water Distribution Systems Analysis Symp., Engineers Australia, Adelaide, Australia.