Jul 13th, 3:50 PM - 4:10 PM

Water Supply System Classification for Water Quality Improvement

Bruno Melo Brentan
LHC - FEC, University of Campinas, brunocivil08@gmail.com

Gustavo Meirelles Lima
LHC - FEC, University of Campinas, limameirelles@gmail.com

Edevar Luvizotto Junior
LHC - FEC, University of Campinas, edevar@fec.unicamp.br

Joaquín Izquierdo
FluIng-IMM Universitat Politècnica de València, jizquier@upv.es

Rafael Pérez-García
FluIng-IMM Universitat Politècnica de València, rperez@upv.es

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

Water Supply System Classification for Water Quality Improvement

Bruno Melo Brentan[a], Gustavo Meirelles Lima[a], Edevar Luvizotto Junior[a], Joaquín Izquierdo[b], Rafael Pérez-García[b]
[a] LHC - FEC, University of Campinas, Campinas, Brazil
[b]FluIng-IMM Universitat Politècnica de València, Valencia, Spain
e-mail: brunocivil08@gmail.com, limameirelles@gmail.com, edevar@fec.unicamp.br, jizquier@upv.es, rperez@upv.es

Abstract: Universal access to drinkable water is a constitutional right guaranteed in Brazil. However, not all cities in this country are able to supply the population with the expected quality. Important actions should be taken to improve the supply of drinkable water. To define a strategic plan for this purpose, classification tools can facilitate the design of plans, by grouping cities with similar quality conditions. During the last decades, neural network approaches have been used in environmental models, allowing more accurate representation of some complex systems. This work proposes the use of self-organizing maps (SOM's) coupled with the *k*-means algorithm to determine city groups (clusters) based on water quality features available at the National System of Sanitation Information (SNIS). Using the Calinski-Harabaz (CH) index for clustering performance analysis, an optimal number of clusters is defined. The objective of this clustering is to clarify the real conditions, to understand the main service deficits from the water quality perspective, and to plan suitable strategies to reduce these deficits.

***Keywords**:* water supply systems; water quality analysis; self-organizing maps; *k*-means clustering

1 INTRODUCTION

The need to cope with population growth coupled with water scarcity and increased production costs propels water supply utilities towards improved efficiency. More often than not, the improvement of existing infrastructures fails to match this accelerated demand growth, causing serious operational problems. Moreover, the lack of effluent treatment, mainly in developing countries, worsens even more the available water resources conditions, causing problems to water treatment. Efficient water management is an important task to avoid these problems. A well designed plan can identify future issues and provide technical and financial resources to ensure quality in water and wastewater services. However, Tupper and Resende (2004) conducted a study in Brazil for the period of 1996-2000, finding that the efficiency of water and wastewater services in some regions vary significantly due to climatic, social and economic impacts, showing that water management still needs major advances.

In Brazil, the Law 11445/2007 establishes guidelines for sanitation, with the following principles: universality, integrity, availability, efficiency and economic sustainability, safety, quality and regularity and integration with water resources management. Albuquerque and Ferreira (2012) highlight that, since the creation of this law, the improvement of sanitation indicators is very slow, requiring major investments to achieve its principles.

In this context, system classification mechanisms could help regulatory authorities identify cities that need further investment. In addition, the benchmarking could be determined and used to establish quality standards. Cabrera et al. (2014) and Lima et al. (2015) present different proposals for water supply system classification based on energy consumption of pumping stations. Berg and Lin (2007),

Thanassoulis (2000), and Scaratti et al. (2013) use the Data Envelopment Analysis (DEA) to classify some cities of Peru, United Kingdom and Brazil, respectively, in relation to management of sanitation services.
The application of SOM's as a clustering tool for database operation (Kohonen et al., 2000) or as early data labelling for the application of classification tools (Izquierdo et al., 2016, Aksela, et al. 2009) has improved in recent years, especially fostered by the availability of large databases and the rapid increase in computer processing.

This paper proposes a classification of Brazilian cities according to the quality of treated water and the compliance level of water and sewage services, key features of population health. Systems with similar characteristics are grouped together through a set of SOMs coupled with the *k-means* algorithm. As a result, the groups are classified, allowing the identification of the cities which need more investment. We claim this classification may be an important tool for government policies.

2 METHOD

2.1 Indicator selection

The features used to cluster the water supply systems are selected from the database of the National System of Sanitation Information (SNIS). Due to the large number of cities that do not have a wastewater system, the indicators related to this service are not used. Instead, they are included in the attribute that simply measures the existence or not of this service, which has a huge impact in people health. Regarding potable water, the five indicators in Table 1 are used as clustering features.

Table 1. Selected indicators

Indicator	Description	Indicator	Description
-	Have a wastewater service?	IN079_AE	Sample conformity - residual chlorine (%)
IN055_AE	Water service index (%)	IN080_AE	Sample conformity - turbidity (%)
IN057_AE	Water fluoridation index (%)	IN085_AE	Sample conformity - total coliforms (%)

The most recent available data, used in this work, is from 2014. Cities with missing information or wrong values for the indicators were disregarded. As a result, 2231 cities are considered for the study, which represents 40% of total number of cities in Brazil.

2.2 Self-organizing maps – SOM's

Different sensory inputs activate different regions of the brain. It occurs primarily by the brain's ability to map input signals (motor, visual, audible, etc.) and to distribute them in an orderly fashion, activating only the responsible parts for responding to this stimulation. Kohonen (1990) suggested a general model of brain mapping, only considering fundamental characteristics of natural behaviour, but organizing the problem for computational processing. From a mathematical viewpoint, the objective of self-organizing maps is to process input data of arbitrary dimension and bring them to one or two-dimensional spaces, using transforming operations to ensure topological similarity (Haykin, 2001).

In general, the algorithm deploys a mesh of neurons in the search space and, along the iteration progress, the mesh is adjusted until the synaptic weights are able to represent the feature space.
The algorithm starts by initializing the synaptic weights of the network. Such initialization can be done randomly, distributing weights in the space, or by an orderly procedure, using squared or hexagonal meshes. Each input vector $\boldsymbol{x}$ with *m* dimensions can be written as

$$\boldsymbol{x} = [x_1, x_2 \dots x_m]^{\mathrm{T}}, \tag{1}$$

and an input weight $\boldsymbol{w}_j$ as

$$\boldsymbol{w}_j = [w_{j1}, w_{j2} \ldots w_{jm}]^T \quad j = 1,2 \ldots l, \tag{2}$$

where l is the number of neurons in the network. The search for best similarity between a weight vector $\boldsymbol{w}_j$ and a default input $\boldsymbol{x}$ can be written as the minimization of the norm of the difference between the two vectors. The neuron, $\boldsymbol{i}(\boldsymbol{x})$, with the greatest similarity with the input vector $\boldsymbol{x}$ can be written as

$$\boldsymbol{i}(\boldsymbol{x}) = \mathbf{argmin}_j \|\boldsymbol{x} - \boldsymbol{w}_j\| \quad j=1, 2, \ldots l. \tag{3}$$

The neuron that satisfies the optimal condition is declared the winner neuron and a topological neighbourhood is associated with it. This neighbourhood defines the activation zone, as suggested by the biological inspiration. Taking into account this biological inspiration, cerebral activation nearby a stimulated neuron is greater than the activation at further distant. In this way, the neighbourhood is defined as an activation region, stimulated by the winner neuron. Figure (1) shows a two-dimensional self-organizing map with an input vector. The darkest circle is the winner neuron and a gray scale shows the influence on the neighbourhood in the adaptive process.

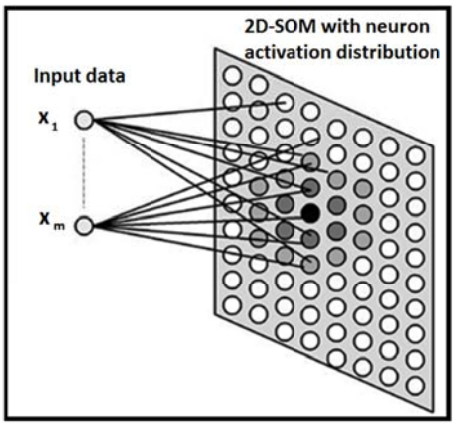


Figure 1. Example of a two-dimensional self-organizing map (adapted from Koua e Kraak, 2004)

Comparing with the brain's behaviour, the activation of the neighbourhood decreases monotonically around the winner neuron. This behaviour can be translated into a mathematical model using a monotonically decaying function. A reasonable choice, usual in the literature, is to take the neighbourhood function as a Gaussian function, as shown in (4):

$$h_{j,i(x)} = exp\left(-\frac{d_{j,i}^2}{2\sigma^2}\right), \tag{4}$$

where $h_{j,i(x)}$ is the topological neighbourhood centred on the winner neuron i, containing the set of j neurons excited by such a winner; σ is the neighborhood size, which can be defined as a decreasing exponential function with respect to the time step. Distance $d_{j,i}$ can be written as the quadratic norm, as shown in (5):

$$d_{j,i} = \|\boldsymbol{r}_j - \boldsymbol{r}_i\|^2, \tag{5}$$

with $\boldsymbol{r}_j$ the position of the excited neuron j and $\boldsymbol{r}_i$ the position of the winner neuron i, measured on the map output space.

Once defined the neighbourhood, each weight is updated considering the entire topological proximity information. The increment $\boldsymbol{w}_j$ is defined by (6):

$$\Delta \boldsymbol{w}_j = \eta h_{j,i(x)} (\boldsymbol{x} - \boldsymbol{w}_j), \tag{6}$$

where η is the forgetfulness rate, which represents the human-like learning process.

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

$$\boldsymbol{w}_j(n+1) = \boldsymbol{w}_j(n) + \eta(n) h_{j,i(x)} (\boldsymbol{x} - \boldsymbol{w}_j(n)). \quad (7)$$

When the learning process is finished, each neuron will be close to an input data group represented in the output space. However, for small group applications, some other procedure is required, because, usually, the number of neurons is higher than the intended number of group. To reduce the cluster number, allowing pre-clustering via SOMs, the *k*-means algorithm is applied.

2.3 *k-means* Algorithm

The *k*-means algorithm has been widely applied as a clustering data method, mainly by it easy application and ability of separation (Herrera et al., 2010, Godin et al., 2005, Laerhofen, 2001. In *k*-means, the number of groups should be previously defined as *k* groups and so the centroid position for each group can be randomly defined inside the input space.
If data is made of *m*-dimensional vectors, as presented in (1), the centroid of group K is written as:

$$\boldsymbol{c}_k = [c_{k1}, c_{k2} \dots c_{km}]^{\mathrm{T}}. \quad (8)$$

For each input data, the distance to all centroids is calculated. An input data is assigned to a group according to the lowest distance to the corresponding centroid. Once defined the cluster for each data, the positions of all the centroids are recalculated, by updating using the average position of the corresponding group. This position may, accordingly, be written as:

$$\boldsymbol{c}_k = \frac{\sum_{i=1}^{n_k} x_i}{n_k}, \quad (9)$$

where n_k is the number of elements belonging to group k.

The objective of the method is to minimize the sum of differences between each input data and the corresponding centroid. After completion, the final position of a centroid corresponds to the average position of its cluster and, eventually, all the input data can be labelled.

The definition of the group number *k* is not always simple, because it depends mainly of the data distribution and its correlations. Furthermore, different to the treatment with labelled data, which allows the use of statistical error measurements, the evaluation of clustering quality requires some specific approach, as presented by Maulik and Bandyopadhyay (2002). Clustering evaluation should take into account not only inter-cluster distances, showing the capacity of the clustering to separate elements, but also intra-cluster distances, which show the capacity to get together around the centroid most elements. The work of Maulik and Bandyopadhyay (2002) presents a set of validation and quality indicator indexes. Among these indexes an important quality indicator of clustering quality is proposed by Calinski and Harabaz (1974) called the CH index, written as:

$$CH = \left[\frac{\sum_{k=1}^{K} n_k \|\boldsymbol{c}_k - \boldsymbol{c}\|^2}{K-1}\right] \Big/ \left[\sum_{k=1}^{K}\sum_{i=1}^{n_k} \frac{\|x_i - \boldsymbol{c}_k\|^2}{n-K}\right] \quad (10)$$

where n_k is the number of elements of cluster k, $\boldsymbol{c}$ 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 variance method (ANOVA) and calculates centroid distances and distances between data and centroids, correlating internal and external distances. This index allows defining a best K partition of data, the highest CH values being linked with the best number of clusters.

3 RESULTS

Once defined the cluster features, a SOM algorithm developed in MatLab was applied for previous clustering. A bi-dimensional network with 20x20 neurons and hexagonal configuration is applied. Figure (2) shows the data maps, corresponding to the distribution of input data and the distance

between weight positions after training. Light colour means short distance, while dark colour means large distance. The input data map analysis is a good tool to understand feature correlations and has been applied for environmental problems recently (Chea et al., 2016, Kalteh et al., 2008). Input maps with similar colour distribution or reverse similar distribution present strong correlation between features. The main correlation observed in Figure (2) is between the presence of wastewater service and water fluoridation. This index is higher for cities without wastewater services. No other correlation may be observed by using only map analysis. Since the SOMs for this data show low correlation between features, the dataset has a very low level of information redundancy. As a result, the quality of the maps describing the dataset may be considered as very good.

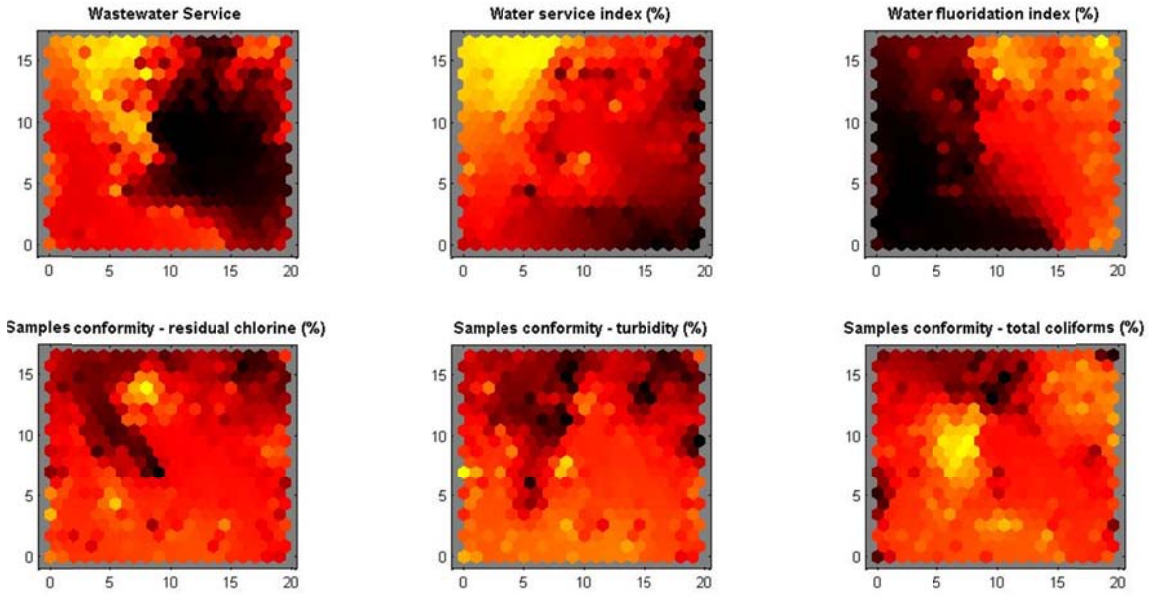
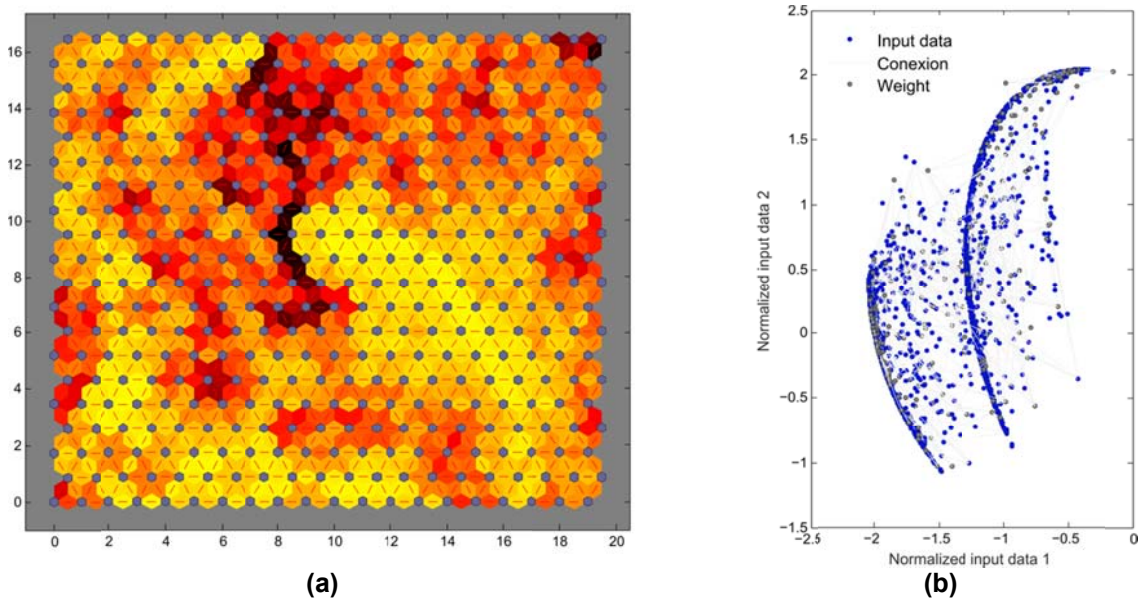


Figure 2. Weight map of each indicator



(a) **(b)**

Figure 3. a) *U* matrix; b) Neuron distribution

Once the feature correlations are defined, an important tool to interpret the SOM results is the *U*-matrix, which presents the final distance at each neuron and its neighbourhood. Following the presented colour scheme, Figure (3-a) shows the *U*-matrix for the performed clustering of water quality. Thus, it is possible to identify boundary regions and macro clusters if a large number of neurons are used, as in this work. Figure (3-b) shows the spatial distribution of neurons and data, in a two-dimensional space, using the first features to represent them. Two macro groups are identified in

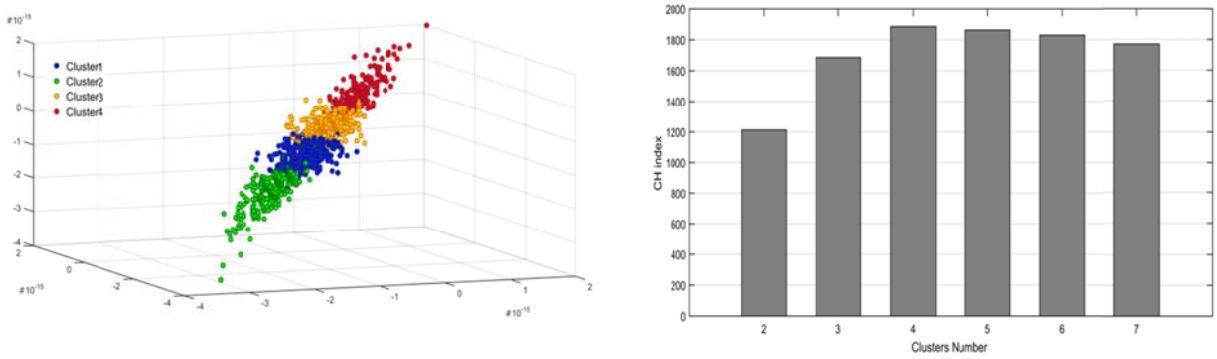
Figure (3-b), and by interpreting the *U*-matrix, also two lighter colour regions separated by a darker region. Lighter regions represent concentration of weights, also corresponding to data concentration. The clustering could have been done with fewer neurons, making the clustering process easier; however, a low number of neurons can harm the quality of data representation because of the dimensionality reduction.

Bearing in mind the main objective of this work, namely, to propose strategic groups for water quality analysis, a restricted number of clusters is need to define clearly actions to improve the performance of water supply utilities. The *k*-means algorithm is applied here to reduce the number of clusters, using the *CH* index to define the ideal number of clusters.

Knowing the weight and input data position, as presented at figure 3-b, it is possible to determine a distance between each data and each neuron. The matrix where all distances are located is noted as dissimilarity matrix. The Euclidian distance is used, considering the continuous distribution of the data in the feature space. For other distribution characteristics, more appropriate ways of calculating the distance may be chosen. The dissimilarity matrix is defined in (11).

$$D = \begin{bmatrix} \|w_1 - x_1\|^2 & \dots & \|w_1 - x_i\|^2 \\ \vdots & \ddots & \vdots \\ \|w_l - x_1\|^2 & \dots & \|w_l - x_i\|^2 \end{bmatrix} \tag{11}$$

Starting with only 2 groups, a *k*-means cluster model with dissimilarity matrix is applied and after classification data, the *CH* index is calculated to evaluate the performance of the clustering. The process is repeated, increasing the number of clusters to find the optimal value. Figure (4) presents the evolution of the *CH* index for the evaluated cluster numbers. Using four clusters presents the best efficiency performance and it is eventually used to divide the data. Figure (5) shows the obtained clusters. The four clusters can be visualized with low superposition of data. The tri-dimensional representation of clusters was chosen by identifying the set of features that represent the clusters with lowest superposition caused by dimensionality reduction.



b)

Figure 4. a) Final clustering after *k*-means processing with 4 groups. b) *CH* index for different number of clusters

Observing the average values of indicators for each group of the final clustering arrangement, as shown in Table 2, well defined characteristics can be identified. Clearly, Group 3 has the best values for all indicators. Group 3 is the only one where most of the cities have a wastewater service, which confirms the concern with water quality. Unfortunately, the number of cities of this group is the small. Group 4 is the second best, closely followed by Group 1. The cities of these groups are characterized by good results in water quality parameters, except for the fluoridation index, but need improvements to universalize water and wastewater services. Group 2 shows a critical situation, where major investments are necessary to achieve universality and quality, which guarantee population health. Group 2 also concentrates most of the cities used for this study, 43.2 % of total, which indicates that Brazil still has a long way to achieve the principles of Law 11445/2007 in its territory.

Table 2. Average values of indicators for each group

Groups	1	2	3	4
Have a wastewater service?	0.4243	0.3797	0.5158	0.4248
Water service index (%)	67.2	64.4	72.0	69.1
Water fluoridation index (%)	45.1	39.1	66.5	52.4
Samples conformity - residual chlorine (%)	74.8	70.9	85.3	81.2
Samples conformity - turbidity (%)	73.9	69.2	83.7	78.1
Samples conformity - total coliforms (%)	77.4	73.9	87.3	82.2
Total cities	634	1006	285	306

To achieve universality in sanitation, a scale factor must be considered. According to Shih et al (2006), an increase of 1% in production reduces unit cost by 0.16%. Therefore, larger systems are more favourable to have good indicators. However, Figure 6 shows no correlation between city population and perceptual water service, water fluoridation and residual chlorine conformity. This behaviour is observed for all indicators, which shows that, besides economic barriers, the adopted policies are fundamental for city sanitation development.

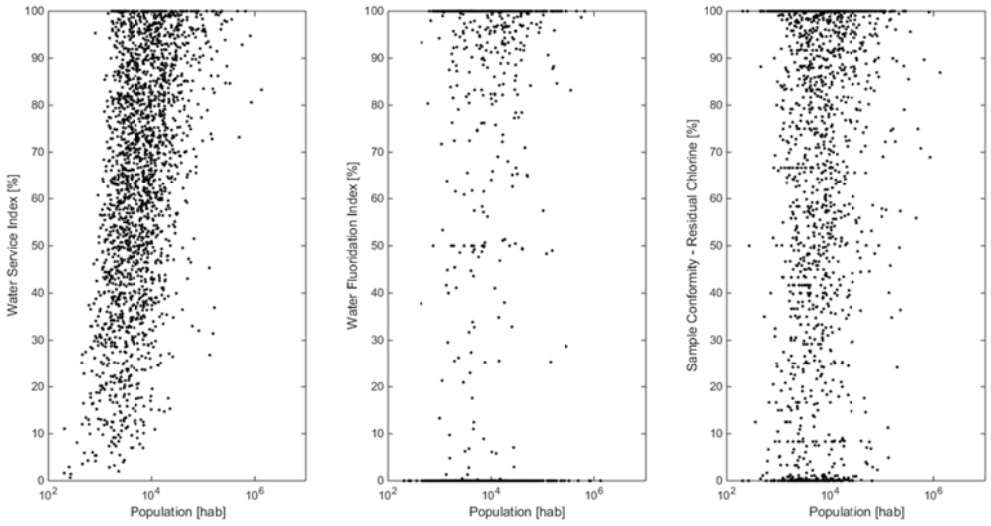


Figure 6. Example of correlation between city size and performance indicators

4 CONCLUSIONS

This paper presents a method to classify cities considering supplied water quality and the existence of a wastewater service. These characteristics are fundamental for population health and are set to achieve the principles established by Law 11445/2007. As observed, most of the cities do not have a wastewater service, which reflects directly into poor water quality. The methodology presented in this work is applied to 40% of Brazilian cities, trying to find strategic groups to improve water quality problem in this country. The clustering analysis showed clearly 4 groups with decreasing quality in the indicators used. The group with worst average values for the indicators concentrated 43.2 % of the total number of cities used in this paper. Considering this study, governmental investments should be addressed, especially for the cities of Group 2, which are very susceptible to epidemic diseases. The

identification of the importance degree of each feature can be a tool to define strategic plans to improve the water quality in Brazil.

REFERENCES

Aksela, K., Aksela, M., Vahala, R., 2009. Leakage detection in a real distribution network using a SOM. Urban Water J., 6(4), 279-289.

Albuquerque, G., Ferreira, A., 2012. O Saneamento Ambiental no Brasil – cenário atual e perspectivas. BNDES 60 Anos Perspectivas Setoriais, Volume 2. Rio de Janeiro.

Berg, S., Lin, C., 2007. Consistency in Performance Rankings: The Peru Water Sector. J. of Appl. Econom., 40(6), 93-805.

Brasil. 2007 Lei nº 11.445, de 5 de janeiro de 2007. Estabelece diretrizes nacionais para o saneamento básico; altera as Leis nos 6.766, de 19 de dezembro de 1979, 8.036, de 11 de maio de 1990, 8.666, de 21 de junho de 1993, 8.937, de 13 de fevereiro de 1995; revoga a Lei nº 6.528, de 11 de maio de 1978; e dá outras providências. Diário Oficial da União, Brasília, DF.

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.

Calínsky, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Comm. Stat. Meth., 3(1), 1-27, 1974.

Chea, R., Gaël G., Sovan, L., 2016. Evidence of Water Quality Degradation in Lower Mekong Basin Revealed by Self-Organizing Map. PlosOne, 11.1.

Godin, N., Huguet, S., Gaertner, R., 2005. Integration of the Kohonen's self-organising map and k-means algorithms for the segmentation of the AE data collected during tensile tests on cross ply composites. NDT & E International, 38(4), 299-309.

Haykin, S. 2001. Neural Networks: A Comprehensive Foundation. Hardcover, 2 edition.

Herrera, M., Canu, S., Karatzoglou, A., Pérez-García, R., Izquierdo J., 2010. An approach to water supply clusters by semi supervised learning. International Environmental Modelling and Software, Modelling for Environment's Sake. Ottawa, Canada.

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.

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(7), 835-845.

Kohonen, T, Kaski, S., Lagus, K. Salojärvi, J., Honkela, J., Paatero, V., Saarela, A., 2000. Self-Organization of a Massive Document Collection. IEEE Trans. Neural Netw., 11(3).

Kohonen, T. 1990. The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

Koua, E. L., Kraak, M., 2004. Geovisualization to support the exploration of larg health and demographic survey data. Int. J. Health Geograph., 3(12).

Laerhoven, K., 2001. Combining the self-organizing map and k-means clustering for on-line classification of sensor data. Artificial Neural Networks – ICANN 2001, 213, 464-469.

Lima, G. M. Viana, A. N. C., Dias Jr., R. S. C., Luvizotto Jr, E. 2015. Classification of water supply systems based on energy efficiency, Water Sci. Tech.: Water Supply, 16(1).

Maulik, U., & Bandyopadhyay, S. 2002. Performance evaluation of some clustering algorithms and validity indices. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(12), 1650-1654.

Scaratti, D., Michelon, W., Scaratti, G., 2013. Avaliação da eficiência da gestão dos serviços municipais de abastecimento de água e esgotamento sanitário utilizando Data Envelopment Analysis. Engenharia Sanitária e Ambiental, 18(4).

Shih, J. S., Harrington, W., Pizer, W. A., Gillingham, K. 2006. Economies of Scale in Community Water Systems. J. Amer. Water Works Assoc., vol. 98, no. 9, pp. 100-108.

Thanassoulis, E., 2000. The use of data envelopment analysis in the regulation of UK water utilities: water distribution. Europ. J. Oper. Res., 126(2), 436-453.

Tupper, H. C., Resende, M., 2004. Efficiency and regulatory issues in the Brazilian water and sewage sector: an empirical study. Utilities Policy, 12(1), 29-40.