Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 119 (2015) 1439 – 1446

Procedia Engineering

www.elsevier.com/locate/procedia

13th Computer Control for Water Industry Conference, CCWI 2015

Exploring patterns in water consumption by clustering

Chrysi Laspidou[a,b,*], Elpiniki Papageorgiou[b,c], Konstantinos Kokkinos[b], Sambit Sahu[d], Arpit Gupta[e], Leandros Tassiulas[f]

[a]*Department of Civil Engineering, University of Thessaly, Pedion Areos, Volos 38334, Greece*
[b]*Information Technologies Institute, Center for Research & Technology Hellas-CERTH, 6th km Charilaou-Thermi Rd., Thermi 57001, Greece*
[c]*Department of Computer Engineering, Technological Educational Institute of Central Greece, 35100 Lamia, Greece*
[d]*IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA*
[e]*Department of Computer Science, Columbia University, New York, NY 10027, USA*
[f]*Department of Electrical Engineering, Yale University, New Haven, CT 06520, USA*

Abstract

Water scarcity, high water demand due to increasing urbanization and the ongoing liberalization of the water and energy markets makes water utilities look into innovative ways to approach consumers, to offer attractive plans and educate them by raising their awareness of their resource use. We analyze water consumption data from a group of consumers at the Greek island of Skiathos, for which we have additional information about features concerning their water consumption patterns. These features are used as input vectors for the construction of Kohonen Self-Organized Maps that are used as classification methods to cluster consumers according to their water consumption. Results show that such analysis can be promising for the automatic classification of water consumers, based on urban water demand data, even if the data is not real-time, or even frequent, since consumptions from standard quarterly water bills are used.

Peer-review under responsibility of the Scientific Committee of CCWI 2015

Keywords: Kohonen Self-Organized Maps (SOM); urban water demand; clustering algorithms; data mining; water consumption analysis

* Corresponding author. Tel.: +30-2421074147; fax: +30-2421074169.
E-mail address: laspidou@iti.gr

Peer-review under responsibility of the Scientific Committee of CCWI 2015
doi:10.1016/j.proeng.2015.08.1004

1. Introduction

Growing population and urbanization, coupled with intense droughts and diminishing availability of freshwater resources due to overexploitation, inadequate management and increased pollution make freshwater security an emerging global issue. Several European countries are affected by water scarcity, while water managers are faced with the pressure to be innovative about how they manage existing water supplies. Water companies are concerned with how to manage the demand for urban water for increasing numbers of businesses and households and have even turned to psychology and behavioral sciences in order to understand individual motivations and behaviors associated with water consumption and conservation [1]. Therefore, understanding how businesses and households use water and whether any unique patterns in this use exist is essential. Information and Communication Technologies (ICT) and social computing can be instrumental in raising awareness of stakeholders on the significance of the water sector in sustainability and can be used to change behaviors and attitudes among citizens [2]. Smart meters have been entering the utility market quickly, while since 2010, smart electricity meters are required by law in Germany (EnWG §21c subpar. 1a,b,c,d), while the water utility market is following this trend as well. Water companies can profit from improved network operations and from matching water supply to water demand on a real-time basis, while the implementation of dynamic water pricing schemes will enable utilities to influence their customer's consumption behavior. The design of rates specific to different consumer segments, as well as the ability to forecast demand can be achieved through the targeted analysis of water consumption data [3]. The ability to predict water demand is at the heart of exploiting ICT to make water use more efficient and to result in cities with smaller water footprints overall. Recent research articles present suitable methodologies, such as fuzzy cognitive maps and neuro-fuzzy inference networks [4, 5] that show the efficiency of predicting water demand including seasonal variability. Furthermore, an important consequence of the liberalization of the water market—in parallel with the electricity market—is the freedom that all customers may choose their water supplier. As new companies enter the market, they will need to approach customers knowing their needs and developing products to suit their preferences, by learning to divide the market and by targeting different segments with different plans and marketing methods [6]. This could lead to the development of personalized contracts defined according to consumption patterns, much like personalized health plans, insurance plans, or mobile phone plans.

Data mining in the energy industry is not uncommon practice nowadays, especially for major consumers [3]. Customer segmentation studies based on consumption data have the potential to reveal characteristic customer load profiles within a diverse and heterogeneous group with very different habits and characteristics. As said above, market liberalization and the potential for acquiring a wealth of data through online real-time smart meters have given rise to data mining approaches that analyze load profile data, via various clustering analysis techniques. This maybe a good time for the water industry to synchronize itself with the electricity and/or gas utility and start analyzing user profiles and employing data mining techniques to characterize water use by segment. Even more importantly, it would be great if these efforts were done in parallel, viewing user profiles and trying to evaluate resource use in a coordinating effort for both water and energy use, since the two uses are intimately related and consumers can be educated to view water and energy use as interconnected. With this work, we investigate whether some of the data mining techniques that have been recently used in the energy sector can be applicable in the water sector as well—a sector that often acts as a "follower" of the developments in the energy sector.

Specifically, in this article, we explore whether consumer type (household or business), or the number of individuals living in a household can be inferred from water consumption profiles. By including in our analysis various features of the consumption profile and using data mining and clustering techniques, we take a first step towards showing whether it is possible to separate groups of customers according to selected properties. Since no real-time data is used, but rather historical time series of ordinary quarterly water billing data for a variety of consumers is analyzed, this is only a first step towards this type of analysis.

2. Materials and Methods

In this article, we analyze municipal water consumption for the highly touristic island of Skiathos, Greece. Skiathos is a typical Mediterranean island with long hot summers and mild winters that experiences a summer water demand as much as six times higher than that in winter, due to intense touristic activity. Tourism results in high

seasonal demand variability, covering the needs of hotels, filling-up swimming pools and an increased number of showers and baths due to the being on vacation, going swimming, hot weather etc. [7, 4]. The island of Skiathos has a small water distribution network with a total of about 3,500 water meters, while the whole town is provided with water by groundwater, via a single drilling. In close collaboration with the Skiathos Water Utility Company, DEYASK[1], we obtained quarterly water-billing data for a group of water consumers that were a mix of households and businesses in the island. It was important to see whether there were any distinct patterns that could be detected taking into account the seasonality of water use of hotels and businesses, as opposed to regular water use by households. As a second step, we conducted a separate analysis for households only to see whether there were any patterns to be detected there.

To investigate the possibility to automatically classify private households and different types of non-residential customers using water consumption data, we performed a preliminary study using water consumption data from 168 water consumers. Our set consists of water consumption coming from customer bills issued once every three months, on a quarterly basis, for over seven years. In total, 30 water consumption data points are available for each water consumer covering the time period from 2006 to 2013. Other information about each consumer is also available, including the type of business (if it is not a household) and the number of occupants, in the case of households. Some generic information about the type of household mostly common in Skiathos is also available by DEYASK, but the dataset could definitely improve with ground-truth information, such as household size, type of water heater in the house (electric or solar), dishwasher availability, existence of large garden or pool, etc. Occupancy is an important variable as well and is worth exploring. Many households in Skiathos are only occupied during the summer—for 3 to 6 months—since many houses are summer homes. Another important variable that is worth exploring in terms of water use is room rental, since many householders choose to rent out a few rooms to tourists.

To better define the problem, it is important to define the classes, i.e. the properties of the water consumers that we expect to be able to infer from the data. To define these classes, we either specify the intended use of the consumer classification scheme, or we base this decision on the data that is actually available. These two approaches have been used in the field of electricity data and are referred to as the *application-driven* and *data-driven* class definition methods, respectively [8]. In this paper, we first follow an elementary *application-driven* methodology and simply define the type of consumer (household or non-residential), then specify the type of non-residential consumer and explore whether the clustering methodology is able to support such classification. We obtained data from the local water utility, DEYASK, on the consumptions of **168** water meters with **3** different consumer types (in the parenthesis after each consumer type, we list the number consumers used in our analysis). We refer to this case study as “MIXED”:

- Households (83);
- Hotels (21);
- Taverns, cafes, fast-food premises and miscellaneous shops—all business types that are usually small and stay open throughout the day and night were all grouped under a single category (64).

Next, we continue with a *data-driven* methodology and analyze the data we have for only one consumer category: householder. Specifically, we extend our analysis to a larger set of **454** households, for which DEYASK has provided us with additional information; this data is the number of occupants per household, so we investigate the potential to automatically classify consumers based on this information. We refer to this case study as “HOUSEHOLDS”. Needless to say that water consumption is already strongly correlated with the number of occupants per household. We confirm that by conducting a simple correlation analysis, which shows that these two quantities are highly correlated. So, with this type of analysis we basically want to investigate if our clustering technique classifies and differentiates a highly correlated data set in different clusters, as we would expect.

[1] http://www.deyaskiathos.gr/

To conduct the investigation of this classification potential, we use the Kohonen Self-Organizing Maps (SOMs) methodology. A SOM allows the representation of a variety of high-dimensional data items, by translating them in a quantitative two-dimensional image in an orderly fashion [9]. Each data point is mapped in a single point in the map, the node, and similarities among the items are indicated by their distances on the map. A SOM is able to cluster data, while at the same time it orders the clusters, forming reduced abstractions of complex data. SOMs are used for analysis in economy to compare enterprises at different levels, in industry to monitor and describe the masses of different input states by ordered clusters, while in science and technology the applications are vast with research objects being classified on the basis of their properties. The method has been in practical use in scientific fields, as well as in finance and other areas since 1982 [10]. It is a special type of artificial neural network that conducts a nonlinear mapping from multi-dimensional data to a low-dimensional space, while it preserves the most important metric and topological relationships of the original data; this way, a SOM actually clusters the data. A SOM relies on unsupervised learning to group input vectors into regions of a map, while each vector is assigned to a specific region depending on its Euclidean distance to already mapped vectors. Clustering procedures are then applied to segment vectors within neighboring regions into clusters [8]. SOMs can also be useful for anomaly detection and missing data reconstruction [12], two functions that can be very useful in the context of this study and smart water in general.

For our analysis, our goal was to provide consumer properties—specifically, household or not and type of non-residential customer—that could possibly be automatically detected through the analysis of water consumption data. These properties correspond to the classes that will be included in the first prototype of our water customer classification system. Therefore, this analysis allows us to check beforehand for the existence of a distinct pattern in the consumption of water consumers that are interesting for our specific case study and that we have the data for. Input vectors consist of features we extract from water consumption data. If a large number of water consumers have the same value for a specific property are mapped to the same cluster on the map, we conclude that it is also possible to classify households according to this property using water data only. We then use SOM to explore the data set and to discover which classes are more meaningful to be included in an automatic classification system. The features we identified as important are all based on consumption values and are shown in Table 1.

Table 1. Features used in the input vectors of the SOM.

Features	Variable	Units
Mean quarterly consumption of all available data points	Qmean	m^3
Ratio of Mean Q over Maximum Q	Qmean/Qmax	-
Ratio of Mean Q for 1st quarter of each year over Mean Q for 3rd quarter of each year	Q1mean/Q3mean	-
Ratio of Mean Q for 2nd quarter of each year over Mean Q for 3rd quarter of each year	Q2mean/Q3mean	-
Ratio of Mean Q for 4th quarter of each year over Mean Q for 3rd quarter of each year	Q4mean/Q3mean	-

Consumption figures correspond to simple averages of the actual consumption billed for each water consumer. A series of ratios are also considered for the input vectors. Ratios are quotients of average consumption values for different periods of the year; specifically, we use ratios with maximum values, as well as with consumptions of the 3rd quarter, since this quarter corresponds to summer, which is significant in Skiathos in terms of water use. All values of the features are normalized using the unit variance scaling method, a necessary normalization since the computed features are used as the components of the input vectors of a SOM [10]. SOM implementation was done using the basic version of the SOM included in the special neural network clustering tool (nnstart) in Matlab [11]. Several commercial software packages and freeware on the SOM are available (see for example SOM Toolbox Team [13]). The SOM in Neural Network Toolbox in Matlab [11] for clustering purpose contains auxiliary analytical procedures and makes use of the MATLAB functions. This SOM toolbox is provided with good and

versatile graphics as well as thoroughly proven statistical analysis programs of the results. Using this Neural Network Clustering Tool, the basic version of the SOM is applied, on which the majority of applications is based.

SOM neurons are connected to adjacent neurons by a neighborhood relationship that defines the map structure. In the SOM Toolbox, topology is divided in local lattice structure and global map shape. The SOM training algorithm moves the weight vectors so that they can span across all the data; this way the map gets organized with neighboring neurons on the grid corresponding to similar weight vectors. According to the toolbox requirements, we used the parameters shown in Table 2 to train the SOM, which was done in batch training [10]. Once the map is constructed, the Tool produces the U-matrix to show the results: the U-matrix visualizes distances between neighboring map units and thus shows the cluster structure of the map. High values of the U-matrix indicate the cluster border, while uniform areas of similar color indicate the clusters themselves. Additional visualization tools include component planes that show the values of one variable in each map unit. Several runs were performed to decide which map size is appropriate for each case study. After mapping the number of "hits" per node, we made sure that there were no empty nodes, or node with very few hits. If this were the case, we tested a smaller map size, making sure that the map was not too small to allow clustering. Resulting map sizes are shown in Table 2.

Table 2. SOM training parameters.

Local map lattice	Global map shape	Map size	Neighborhood radius	Initial learning rate	Epochs
Hexagonal	Sheet	7x7 (MIXED) 6x6 (HOUSEHOLDS)	$\sigma_{fin} = 1$	$\alpha_0 = 0.95$	2,000

3. Results and Discussion

We used the SOM Matlab Toolbox to determine the final number of clusters into which the input data is grouped. The tool specifically determines a first set of regions on the map using all the input vectors and then applies a k-clustering filter to balance the map and reduce the total number of regions. In Figure 1, we show the two maps obtained for the two case studies—MIXED and HOUSEHOLDS—from providing the SOM with input vectors that contain all the features described in Table 1. The input vectors are grouped in 4 and 3 clusters for the two case studies, respectively, as shown in Figure 1. Some clusters are larger in size than others, while some contain fewer hits than others—the number of hits for each cluster is shown in Figure 1.

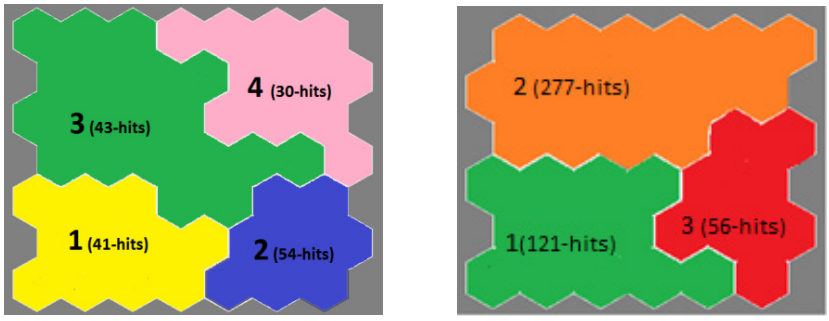


Fig. 1. Clusters obtained with SOM analysis, using input vectors with Table 1 features: (a) "MIXED" case study, 7x7 map, 4 clusters and corresponding number of hits;.(b) "HOUSEHOLD" case study, 6x6 map, 3 clusters and corresponding number of hits.

In order to draw the conclusions that we are actually interested in, i.e., if the features computed on the water consumption data actually cause consumers with similar properties to get "naturally" grouped together on the map, we analyze the data as follows: For the "MIXED" case study, we graph in Fig. 2 the percentage of water consumers that are residential and non-residential in each of the clusters identified by the SOM. The plot in Fig. 2a shows that over 60% of residential properties are grouped under Cluster 3, while this percentage drops consistently for each cluster, while the opposite is true for non-residential consumers which increase steadily in percentage from Cluster 3

and through the rest of the clusters. Such clustering is a positive sign that the property "residential" or "non-residential" can be distinguished according to the clustering provided by the SOM. Further analysis in the residential or non-residential properties that group under different clusters would be necessary, in order to discover which special feature classifies each property in each cluster. In Fig. 2b, we show a similar analysis for non-residential properties only: we see that hotels seem to differentiate themselves among the non-residential properties, showing a clear trend among the different clusters. Again, this is a good sign that using the set of features that we have used here is likely to be useful in determining where each property belongs. The analysis becomes more and more interesting the more data is collected (such as family income, house surface area, number of bedrooms, garden, swimming pool, solar or electric water heater, summer home or permanent residence, etc.) so the method has a great potential and could be developed further with success.

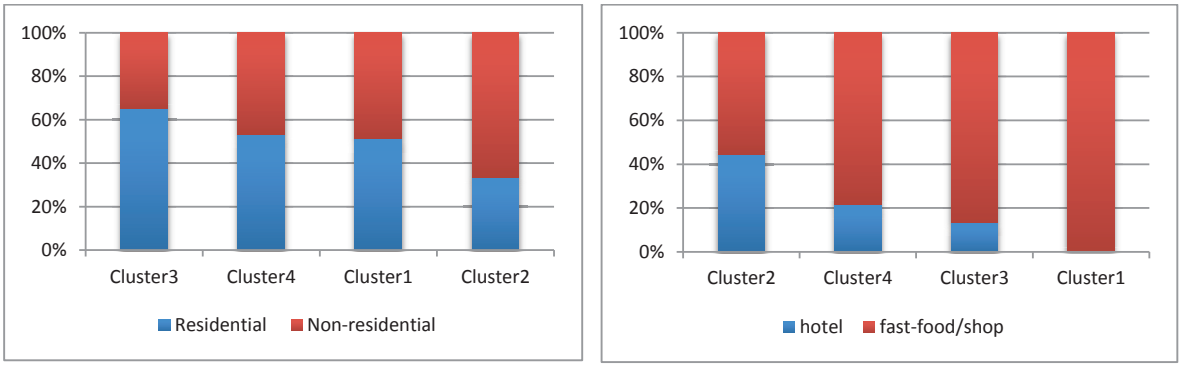


Fig. 2. "MIXED" case study: (a) Distribution of residential and non-residential properties among clusters; (b) distribution of different business types among clusters.

Similar kind of analysis is conducted for the second case study, namely the "HOUSEHOLD". The feature that we study here is the number of occupants per household. We group these features in three categories: the first includes 1 occupant per household, the second includes 2 to 4 occupants and the last one includes 5 occupants per household or more. The graph of this analysis is shown in Fig. 3(a) and the results are similar to those observed for the "MIXED"

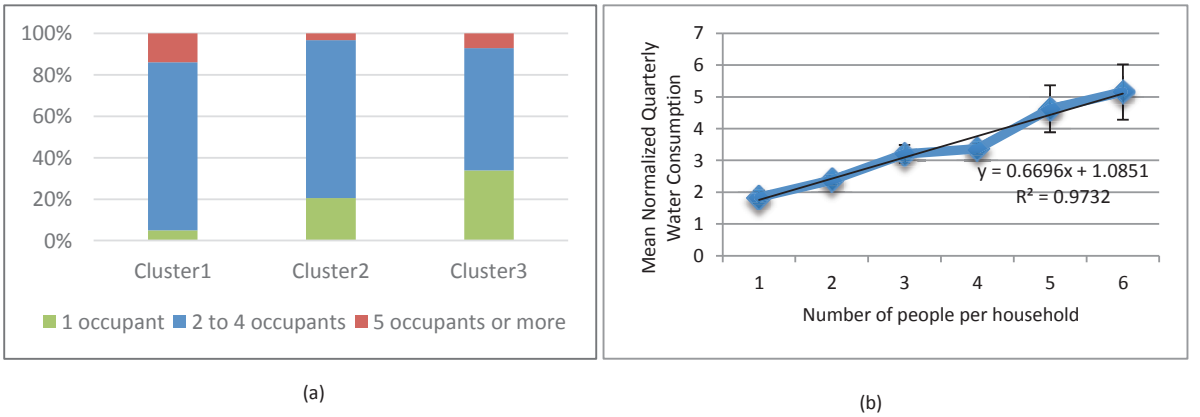
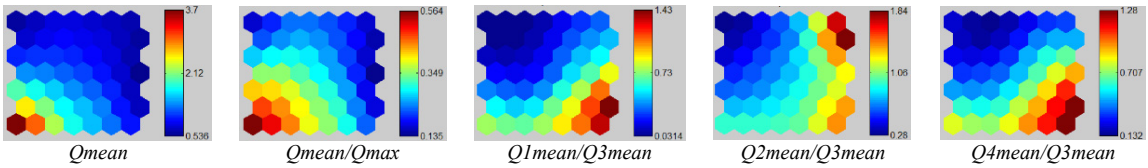


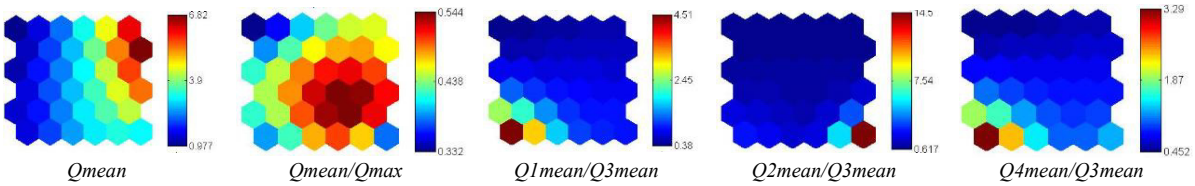
Fig. 3. "HOUSEHOLD" case study: (a) Cluster analysis and (b) plot and linear regression of water consumption vs. number of occupants in household.

case study. We see that the two initial categories ("1 occupant" and "2 to 4 occupants") follow a clear trend among the clusters. The third category, "5 or more occupants" does not have such a clear trend, but it is only a small percentage of the total and represents probably not a typical Greek household (due to the large number of occupants). In this case study, the pattern we are trying to "discover" is not really difficult to detect, since the data is by itself directly related to the number of occupants in the household. This can be seen in Figure 3(b), in which we show a plot of water consumption vs. number of occupants in a household. The two variables are linearly correlated, with linear regression giving a R^2 of almost 1.0. The SOM manages to capture this trend, although it seems to be a tool mostly suited to map multi-dimensional complex data and detect patterns in them.

To further analyze the clustering performance of the SOM, it is interesting to observe the *component planes* that come out of the SOM analysis. A component plane displays how a single feature is distributed over the map or, equivalently, how the feature contributes to the final shape of the map. The final SOM map is a combination of these layers that come from each variable considered in the input vectors. Component planes for both case studies are shown in Figure 4. They show, for example, that consumers with low average consumption (Qmean) are mapped on the top and right outer edges of the map, while high consumptions are mapped in the left lower corner. The layering of this mapping can be seen in the component plane plots, which can provide a lot of additional information on why properties are included in specific clusters. These component planes need to be studied together with the SOM maps that show the clusters (Fig. 1), which correspond areas of the map in specific clusters.



(a)



(b)

Fig. 4. Component planes (a) "MIXED" case study; (b) "HOUSEHOLD" case study

4. Conclusions

In this article, we have conducted a preliminary analysis of urban water consumption data, in order to automatically classify them in different categories, based on their pattern of consumption for the island of Skiathos. We have presented our analysis for two case studies: one includes different water consumers, such as households and local businesses, while the other includes only households with various numbers of occupants. Water consumption data came from billing records issued on a trimester basis. Our results show that such analysis can be promising for the automatic classification that we wanted to achieve, based on urban water demand data, even if the data was not real-time, or even frequent, but were simply consumptions from standard quarterly water bills. When cluster analysis was performed and additional data were provided for the water consumers, we saw that there are properties, such as whether a water consumer is residential or non-residential, that could be inferable from water consumption data. When analyzing only household data, we showed that number of occupants is an important property that could be inferred from this data. The identification of such properties represents a necessary first step towards the investigation of the potential of automatic classification of private households using water consumption

data. Collecting information on more properties that could be linked to water consumption is the next important step.

Acknowledgements

This work was supported by the project SOFON, which is implemented under the "ARISTEIA" Action of the "OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING" and is co-funded by the European Social Fund (ESF) and National Resources. The authors also wish to acknowledge the Director of DEYASK, Mr. I. Sarris for his collaboration in providing us with data for this study.

References

[1] B. Jorgensen, J. Martin, M. Pearce and E. Willis, Predicting Household Water Consumption With Individual-Level Variables, Environment and Behavior, 46(7) (2014) 872-897.
[2] C. Laspidou, ICT and stakeholder participation for improved urban water management in the cities of the future, Water Utility Journal 8 (2014) 79-85.
[3] C. Flath, D. Nicolay, T. Conte, C. van Dinther, L. Filipova-Neumann, Cluster Analysis of Smart Metering Data An implementation in Practice, Business & Information Systems Engineering. 1 (2012) 31–39.
[4] E.I. Papageorgiou, K. Poczeta and C. Laspidou, Application of Fuzzy Cognitive Maps to Water Demand Prediction, Proceedings IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2015, August 2-5, Istanbul, Turkey.
[5] E.I. Papageorgiou, K. Poczeta, A. Yastrebov, and C. Laspidou, Fuzzy Cognitive Maps and Multi-step Gradient Methods for Prediction: Applications to Electricity Consumption and Stock Exchange Returns, in: R. Neves-Silva, L.C. Jain and R.J. Howlett (Eds.) Intelligent Decision Technologies, Springer International Publishing Switzerland 2015, pp. 501-511.
[6] F. Rodrigues, J. Duarte, V. Figueiredo, Z. Vale, M. Cordeiro, A Comparative Analysis of Clustering Algorithms Applied to Load Profiling, in: P. Perner and A. Rosenfeld (Eds.), MLDM 2003, Lecture Notes in Artificial Intelligence 2734, Springer-Verlag Berlin Heidelberg, 2003, pp. 73-85.
[7] D. Kofinas, N. Mellios, E. Papageorgiou, C. Laspidou, Urban Water Demand Forecasting for the Island of Skiathos, Procedia Engineering, 89 (2014) 1023-1030.
[8] C. Beckel, Ch., Sadamori, L., and S. Santini, Towards automatic classification of private households using electricity consumption data. Buildsys'12, November 6, 2012, Toronto, ON, Canada.
[9] T. Kohonen, Essentials of the self-organizing map, Neural Networks, 37 (2013) 52-65.
[10] T. Kohonen, MATLAB Implementations and Applications of the Self-Organizing Map, Unigrafia Oy, Helsinki, 2014.
[11] SOM in Matlab, http://www.mathworks.com/help/nnet/gs/cluster-data-with-a-self-organizing-map.html
[12] B. Lamrini, E.K. Lakhal, L. Wehenkel, Data validation and missing data reconstruction using self-organizing map for water treatment, Neural Computing & Applications, 20(4) (2011) 575-588.
[13] SOM Toolbox Team (1999). http://www.cis.hut.fi/projects/somtoolbox/documentation/. http://www.cis.hut.fi/projects/somtoolbox/package/papers/techrep.pdf.