European Water 58: 443-448, 2017.
© 2017 *E.W. Publications*

Data mining for household water consumption analysis using self-organizing maps

A.E. Ioannou, D. Kofinas, A. Spyropoulou and C. Laspidou*
Department of Civil Engineering, University of Thessaly, Pedion Areos, Volos 38334, Greece
** e-mail: laspidou@uth.gr*

Abstract: Household water consumption is a part of the human related water cycle that can get into the core of water resources management. Analysis of water consumption data can reveal great potentials of individualized water services planning. Data mining is the process of identifying and extracting potentially useful information from data sets. Self-Organizing Maps (SOMs) is a data mining technique that involves an unsupervised learning method to analyze, cluster, and model various types of large data sets. In this paper, it is presented how the daily water consumption of a household in Sosnowiec, Poland, can be clustered into days of the week, through some features. The features used to discretize the days of water consumption are statistic metrics and time zone consumption metrics. The time zoning is realized in two ways, the first being the typical morning, noon, afternoon, evening and night and the second considering the local working hour time zones of three main working sectors, banks, offices and shops. We use the SOM algorithm in three approaches. In each approach, we use some of the selected features. We have managed to get some clusters with specific features that divide the days of this household in weekdays and weekends.

Key words: Data mining, self-organizing maps, water consumption analysis

1. INTRODUCTION

Smart Cities have been defined in many ways. Special focus is given on Information and Communication Technologies (ICTs) as a key driver. Other important aspects of instrumenting the landscape of a smart city would be society, economy, and policies driven by participatory processes. The concept of Smart Cities sets, in the core of urban planning, monitoring, recording, analyzing and transferring in real time all data relative to social activity; thus, enabling social and governmental awareness of all significant design variables. The monitored variables are linked to basic networks and environments such as city services, transport, water and energy (Manville et al., 2014).

Subsequently, the need of collecting and managing great amounts of temporal and spatial high-granularity data becomes of conclusive importance. Great progress in ICTs, social network, and Data Mining (DM) enable new paths of urban planning, empowering the resilience of urban infrastructure and the adequacy of resources and systems (Laspidou, 2014; Yang et al., 2017).

DM is an ongoing research area, responding to the presence of large databases in commerce, industry and research (Bishop, 1995). It is a very useful technique in research; because, not only can someone extract meaningful information from certain databases, but can also predict changes, discover patterns and relations hidden among the data (Fayyad et al., 1996). In addition, DM is an interactive process that requires the intuition, background knowledge and the computational efficiency of modern computer technology. Clustering techniques can reveal patterns and identify similar trends and components linked to similar behaviors. Clustering visualization is a very important part of DM and can be successfully accomplished, by the use of self-organizing maps (Kaski, 1997).

One of the fundamental urban activities, firmly related to all human well-being aspects, such as public health, is the urban water consumption. The Water Distribution Networks (WDNs) are to be thoroughly monitored in different scales, from an aggregated urban water resources view, down to a

household water consumption view, so as to gather all useful data that will enable the improvement of water resources management, through all the available data analysis techniques. The aforementioned DM techniques can realize water consumption forecasting at small scales, by clustering householders in different consumption patterns; thus, enabling water utilities to get more efficient by heading towards more individualized customer services (Beckel et al., 2012; Laspidou et al., 2015). Parallel to increasing resolution at individualized scale, analysis can also enable temporal clustering resolution, identifying different daily consumption patterns. Past research (Arampatzis et al., 2014) has already indicated weekdays/weekends water use particularities at urban level, revealing the potential to further investigate daily patterns at household level.

The aim of the present study is to investigate how specific daily household patterns can be inferred from household water consumption profiles. We make an effort to cluster days of similar water consumption profiles. The automatic classification of days into clusters is accomplished with the application of the SOM methodology.

2. MATERIALS AND METHODS

In this paper we develop a methodology for the detection of daily patterns in household water consumption, through a detailed patterns analysis using SOM (Kohonen, 1995). The methodology is a popular neural network algorithm based on unsupervised learning. SOM uses specifically chosen features of a population; then it calculates the Euclidean distances of each unit of the population considering the features as dimensions or components of the input vectors. Lastly, it converts the multidimensional positions of the units into 2-dimensional and maps them. These maps depict all units in a sense that neighboring units correspond to similar features; thus making clustering of similar units possible. The SOM has proven to be a valuable tool in DM with applications in financial data analysis (Deboeck et al, 1998), in engineering applications, pattern recognition (Kohonen et al., 1996), image analysis, process monitoring and fault diagnosis (Simula et al., 1995). SOM method is widely used to cluster a variety of attributes such as fine-grained electricity consumption, water consumption, etc. (Beckel et al., 2012; Laspidou et al., 2015; Räsänen et al., 2008).

Water consumption is recorded by sensors installed in individual households. Input vectors are produced from water consumption data from a household in Sosnowiec, an industrial city in southern Poland. In this household, sensors were installed in seven different faucets, so we summed all our data in order to have time-series of the total consumption of the house. Water consumption values are recorded every 30 sec and a 445-days sample is used for the analysis. Each day is considered to be a unit of a population of 445 and the analysis aims at clustering days that are linked to similar profiles by mapping them closely.

We use 13 features to describe the daily consumption behavior, listed in *Table 1*. The features refer to total daily consumption (feature 1), daily standard deviation (σ) of daily water consumption (feature 6) and partitioning of daily consumption into specifically chosen time-zones (features 2-5 & 7-13). Our data are divided in the following time zones: morning (feature 2), noon (feature 3), afternoon (feature 4), evening (feature 5) and night (feature 7). Another time zone division considers the working hours and rest hours of offices (feature 8 & 10, respectively), banks (features 9 & 11, respectively) and shops (features 12 & 13, respectively). The features are chosen in the sense that total consumption, σ and time zones consumption are suspected to possibly characterize different days' patters. Specifically the working hours' features are taken into account, because the household is located in an urban environment and most probably its water consumption is directly affected by the working routines. The aforementioned features are combined to create three different dimension sets for the three approaches of our investigation. The features regarded for each approach are listed in *Table 1*.

After the preparation of the data and estimation of the feature table of its single day of consumption, a matlab SOM-toolbox and its scripts are used to preprocess data, initialize, normalize and train SOM and construct the maps (Vesanto et al., 1999). The normalization of data is of great

importance for a very specific reason: The estimated features and components of the input vectors of a SOM, due to their different nature, exhibit values of very different magnitude. The feature with the greater magnitude would affect the Euclidean distance unequally to the features with smaller magnitude. With the normalization we achieve, equal partitioning of the features.

Another very significant pre-process is that of labeling each feature table after its corresponding day. The purpose of this process is that the maps created are legible and comprehensible and the clusters, possibly created, are easily visualized. This means, that we expect Saturdays, for example, to gather around the same area of a SOM map and this way construct a cluster of similar consumption profile corresponding to that of a typical Saturday.

Table 1. *Features of daily consumptions used for the 3 different SOM approaches*

approach 1	**approach 2**	**approach 3**
1. total daily consumption	7. night consumption *(12am-8am)*	8. consumption at offices working hours *(8am-4pm)*
2. morning consumption *(6am-10am)*	8. consumption at offices working hours *(8am-4pm)*	9. consumption at bank working hours *(8am-6pm)*
3. noon consumption *(10am-2pm)*	9. consumption at bank working hours *(8am-6pm)*	12. consumption at shops working hours *(11am-7pm)*
4. afternoon consumption *(2pm-6pm)*	10. consumption at office rest hours *(4pm-10pm)*	
5. evening consumption *(6pm-10am)*	11. consumption at bank rest hours *(6pm-10pm)*	
6. standard deviation of water consumption (σ)	12. consumption at shops working hours *(11am-7pm)*	
	13. consumption at shops rest hours *(7pm-10pm)*	

3. RESULTS AND DISCUSSION

The results of the runs are depicted with use of three kinds of SOM maps (*Figures 1, 2 and 3*). The U-matrix is the 2-dimensional depiction of the multi-dimensional Euclidean distances of all the daily consumption profiles. As we see from top to bottom there is gradation from dark blue, small distances to light yellow, big distances. This means that two units in neighboring cells are closer if the cells are in the blue area than if they were in the yellow area. In the labels map, the location of each unit is presented according to its characterization as a day of the week. This way it can be revealed if same days of the week tend to cluster in neighboring cells; thus prove to have similar water consumption profiles. At the right side of the figures, the feature matrixes are presented. These maps depict the behavior of each feature separately.

In approach 1 (*Figure 1*), we use features 1 to 6, namely, total daily, morning, noon; afternoon, and evening consumptions and σ. Two main cluster areas, area 1 and area 2, are formed. In the first one, weekends are more frequent, while as, in the second one, weekdays seem to dominate. This clustering of weekdays separately than weekends proves that in the household depicts different water use profiles in working days than in non-working. Moreover, weekend cluster cells are much closer one another than weekday cells. That means that weekend profiles are more uniform than the weekdays. In area 2, three sub-clusters 2a, 2b, and 2c seem to be formed. Sub-cluster 2a includes only Tuesdays, sub-cluster 2b includes only Wednesdays and finally sub-cluster 2c includes only Mondays. The sub-clustering of cluster 2 justifies its looseness, as it seems that throughout the working days, some more specific daily consumption prototypes are formed with some differences. Observing the feature matrixes, we can see that total daily consumption and σ are similar. These two show that weekends in the upper are firmly located, in contrast to weekdays which are more loosely located. This pattern is almost followed by the afternoon consumption feature. On the other hand, morning, noon, and evening consumption features are more uniform in an extended area, with some yellow "gorges" low at the weekdays' area. Subsequently the U-matrix is formed as a combination of the six features degrading top to bottom from blue to yellow.

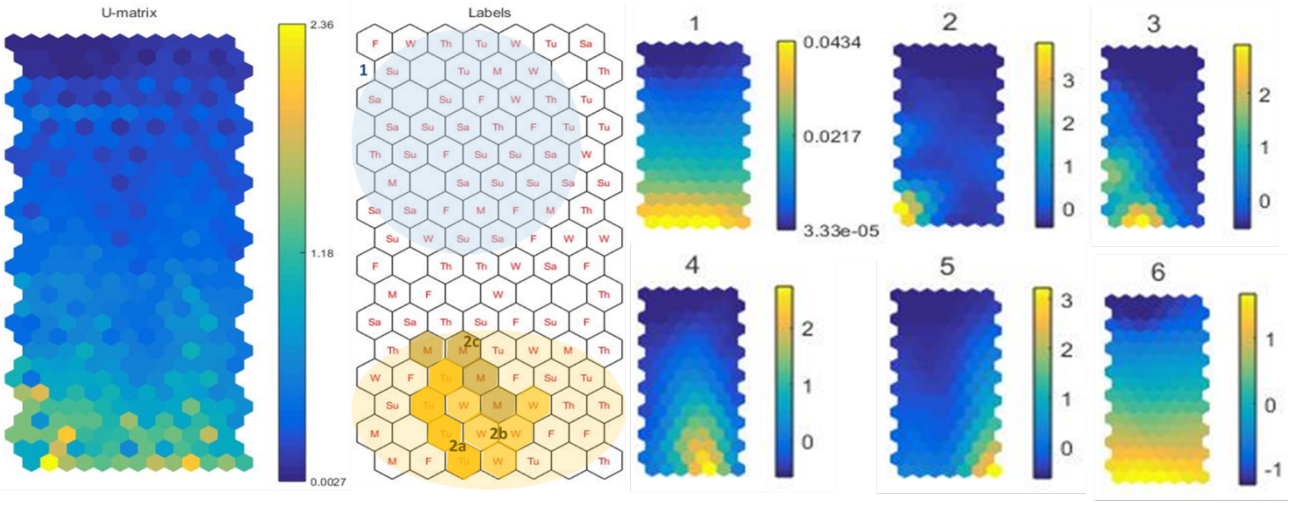

Figure 1. SOM map results: U-matrix, Labels, feature 1 (total daily consumption), feature 2 (morning consumption), feature 3 (noon consumption), feature 4 (afternoon consumption), feature 5 (evening consumption), and feature 6 (σ) (Approach 1).

In Approach 2 (*Figure 2*), we use features 7 to 13, which refer to the working and resting hours consumptions as well as the night consumption. With this approach, we see that clusters are formed in the same way (weekends upper and weekdays down). However, the clusters appear more discretized and more extended than those of approach 1. This proves that the working and resting hours features are more proper than the typical time zoning of morning, noon, afternoon, and evening features. This could be an indicator that the case study household members follow the prevailing working routine of Sosnowiec. In contrary to approach 1, two sub clusters, 1a and 1b, are formed in cluster area 1 rather than in area 2. The cluster 1 mostly contains weekends, while its sub clusters 1a and 1b contain exclusively Saturdays and Sundays. The weekdays in this approach don't seem to further sub-cluster discretized. In Approach 3 (*Figure 3*), only three features are used, the working hours consumptions. This improves clustering, since in the weekend area, Saturdays and Sundays rise up to 3 out of 4. In the weekday area, 86 % of the cells are occupied by weekdays. In all approaches, we can see that Fridays are equally distributed in both areas. This indicates that Fridays water consumption profiles are somewhere in between and have characteristics of both weekday and weekend profiles.

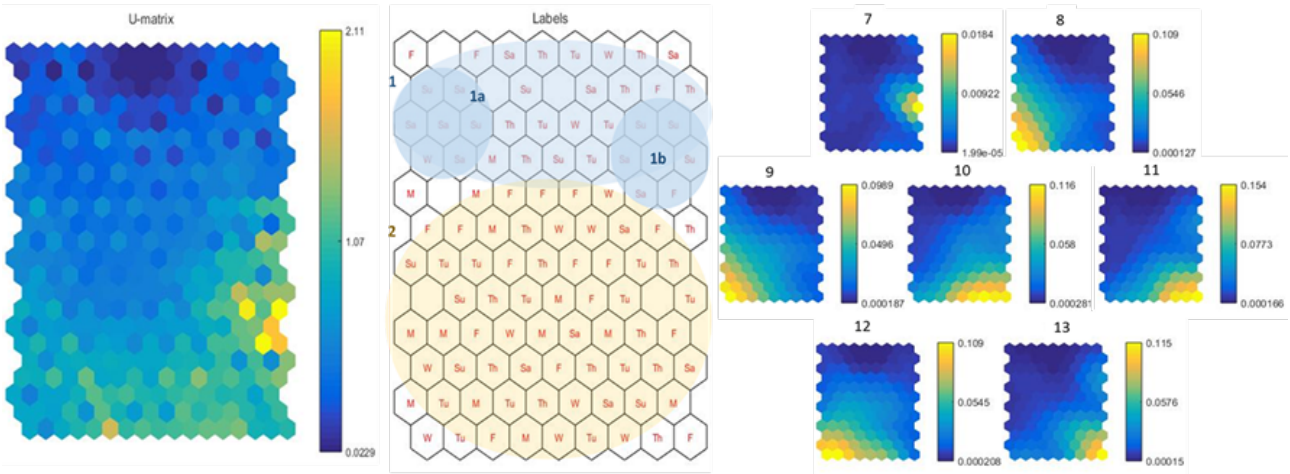

Figure 2. SOM map results: U-matrix, Labels, feature 7 (night consumption), feature 8 (consumption at offices working hours), feature 9 (consumption at banks working hours), feature 10 (consumption at offices rest hours), feature 11 (consumption at banks rest hours), feature 12 (consumption at shops working hours), and feature 13 (consumption at shops rest hours) (Approach 2).

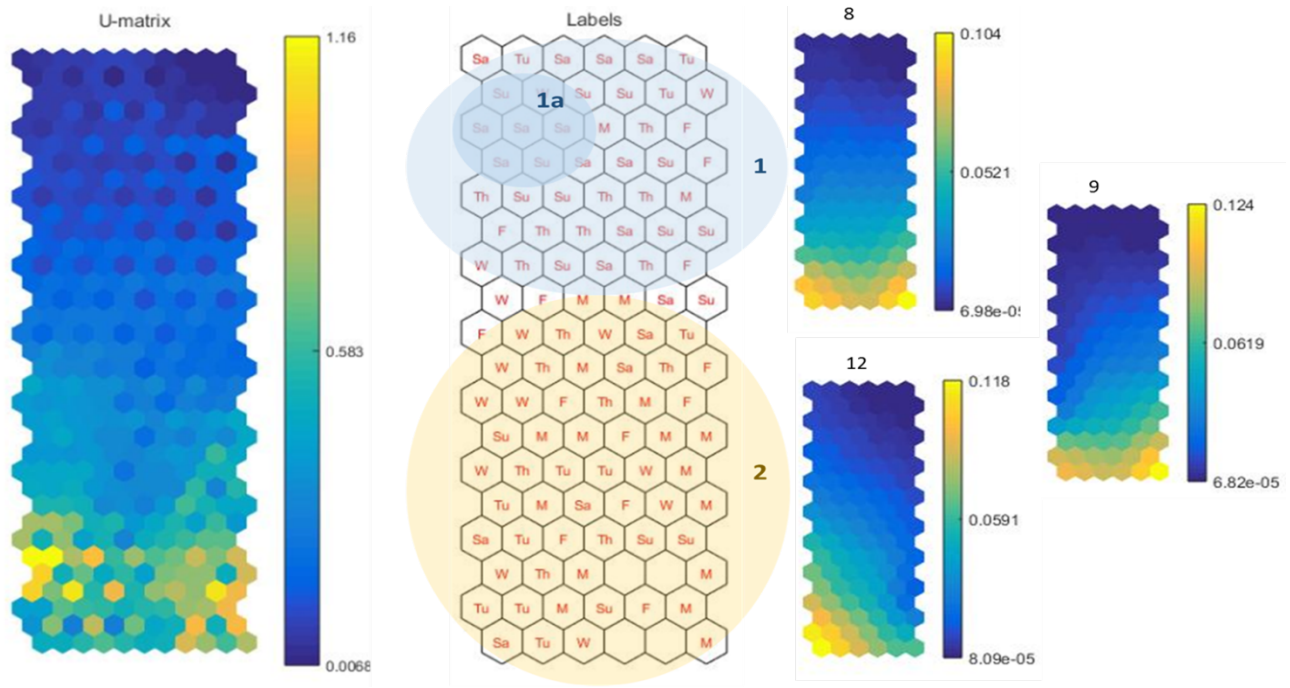

Figure 3. SOM map results: U-matrix, Labels, feature 8 (consumption at offices working hours), feature 9 (consumption at banks working hours), and feature 12 (consumption at shops working hours) (Approach 3).

4. CONCLUSIONS

In this paper, we perform automatic classification of daily water consumption patterns for a household in Sosnowiec, using data collected by sensors. Our investigation is built upon three approaches accomplished with use of the SOM algorithm. Thirteen descriptive features of daily consumption patterns are used for the classification. Approaches 1 and 2, distinguishes two water consumption profiles dividing days of consumption into two main clusters, weekdays and weekends. In approach 3, with use of features partitioning daily water consumption into the time-zones of Sosnowiec working hours, the two clusters get more solid and the methodology gets more efficient. In all approaches, the clustering of days of consumption into weekdays and weekends implies that the household is affected by the urban working routine.

In future scenarios, this investigation could be expanded to more households in order to achieve a further classification, even manage to identify each day of the week. Investigation among households in working areas, or leisure places might indicate different water consumption daily patterns. In all cases the choice of good descriptive features seems to be important. The features have to be customized to better describe daily routines of different types of household.

ACKNOWLEDGMENTS

This work was supported by the project Water4Cities - Holistic Surface Water and Groundwater Management for Sustainable Cities - which is implemented in the framework of the EU Horizon2020 Program, Grant Agreement Number 734409.

REFERENCES

Arampatzis, G., Perdikeas, N., Kampragou, E., Scaloubakas, P., and Assimacopoulos, D., 2014. A water demand forecasting methodology for supporting day-to-day management of water distribution systems. In 12th International Conference “Protection and Restoration of the Environment”, Skiathos, Thessaloniki, Greece.

Beckel, C., Sadamori, L., and Santini, S. 2012. Towards automatic classification of private households using electricity consumption data. In Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings), ACM, pp. 169-176

Bishop, C.M. 1995. Neural Networks for Pattern Recognition, Oxford University Press, Oxford.

Deboeck, G and Kohonen T. 1998. Visual Explorations in Finance using self-organizing Maps, Springer, London.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy R. 1996. Advances in knowledge Discovery and Data Mining, AAAI Press/The MIT Press, California.

Kaski, S. 1997. Data Exploration Using Self-Organizing Maps, Ph.D. thesis, Helsinki University of Technology.

Kohonen, T. 1995. Self-Organizing Maps, Springer, Berlin.

Kohonen, T., Oja, E., Simula, O., Visa, A. and Kangas, J. 1996. Engineering applications of the self-organizing map. Proceedings of the IEEE, 84(10), 1358-1384.

Laspidou, C. 2014. ICT and stakeholder participation for improved urban water management in the cities of the future. Water Util. J, 8, 79-85.

Laspidou, C., Papageorgiou, E., Kokkinos, K., Sahu, S., Gupta, A. and Tassiulas, L. 2015. Exploring patterns in water consumption by clustering. Procedia Engineering,119, 1439-1446.

Räsänen, T., Ruuskanen, J., and Kolehmainen, M. 2008 Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. Applied Energy, 85(9), 830-840.

Manville, C., Cochrane, G., Cave, J., Millard, J., Pederson, J. K., Thaarup, R. K., and Kotterinkf, B. 2014. Mapping smart cities in the EU.

Simula, O. and Kangas, J. 1995. Process monitoring and visualization using self-organizing maps, in: Neural Networks for Chemical Engineers, Computer-Aided Chemical Engineering, vol 6, Elsevier, Amsterdam.

Vesanto, J., Himberg, J., Alhoniemi, E. Parhankangas, J. 1999, November. Self-organizing map in Matlab: the SOM Toolbox. In Proceedings of the Matlab DSP conference, 99, 16-17.

Yang, L., Yang, S.-H., Magiera, E., Froelich, W., Jach, T. and Laspidou, C. 2017. Domestic water consumption monitoring and behaviour intervention by employing the internet of things technologies. Procedia Computer Science, 111, 367-375.