Behavioral Profiles for Building Energy Performance Using eXclusive SOM

Félix Iglesias Vázquez[1], Sergio Cantos Gaceo[2],
Wolfgang Kastner[1], and José A. Montero Morales[2]

[1] Vienna University of Technology, Automation Systems Group,
Treitlstr. 1-3/ 4. Floor, A-1040 Vienna, Austria
{vazquez,k}@auto.tuwien.ac.at
https://www.auto.tuwien.ac.at
[2] La Salle (Universitat Ramon Llull), Electronics & Communication,
Passeig Bonanova 8, 08022 Barcelona, Spain
{scantos,montero}@salle.url.edu
http://www.salle.url.edu/portal/departaments/home-depts-DEC

Abstract. *The identification of user and usage profiles in the built environment is of vital importance both for energy performance analysis and smart control purposes. Clustering tools are a suitable means as they are able to discover representative patterns from a myriad of collected data. In this work, the methodology of an eXclusive Self-Organizing Map (XSOM) is proposed as an evolution of a Kohonen map with outlier rejection capabilities. As will be shown, XSOM characteristics fit perfectly with the targeted application areas.*

Keywords: pattern discovery, neural network, Self-Organizing Map, user behavior and profiling, energy performance simulation, building automation.

1 Introduction

Clustering techniques are usually deployed to discover representative patterns or profiles in diverse fields and applications. In this work, we focus on energy performance of buildings, and smart home and building control. Here reliable data are necessary to calculate and predict (energy) consumption. Besides information of the building structure, layout and physics, and environmental data (e.g. weather data), the identification of energy user models in the built environment is outstanding for energy efficiency purposes [14], but also for the evaluation of energy simulations [13]. Moreover, behavioral profiles (wrt. occupancy, setpoint temperatures, device usage, lighting habits, water usage, domestic hot water, heating, cooling, ventilation or electricity consumption, etc.) will improve and refine the way for a smarter control of sustainable buildings in the daily-life [8].

Self-Organizing Maps (SOM) have been previously used in similar scenarios, like usage patterns (e.g. [12]) or the identification of energy user profiles (e.g. [20]). A well-known problem in clustering is the existence of outlying elements

that disturb the reliability of the results. In this paper, we present eXclusive SOM (XSOM) as a SOM evolution. The main advantage of XSOM is the capability of rejecting and filtering non-representative and erratic data (outliers) that SOM classifies likewise, a problem already detected in previous works [15].

Therefore, the present work introduces the XSOM method and tests it in a behavioral profiling case with outliers, comparing it with SOM and K-means clustering performances. On the other hand, the importance of behavior and usage profiles in the building analysis and control is emphasized, as well as the convenience of using XSOM in this application field.

2 Behaviour Profiles and Uncertainties in their Discovery

In the built environment, energy profiling is related to the analysis of the current or predicted energy performance of buildings. Nowadays, architects, construction engineers, and facility designers value the importance of energy profiling. Still, they are forced to rest their decisions on experienced knowledge or trust in values that often do not represent the reality. Obviously, the result is not as accurate as it should be, specifically in energy behavior issues [14].

Energy behavior describes the way (habits) in which inhabitants use or affect – in a direct or indirect fashion – the diverse energy resources of a dwelling. Modeling the behavior is important as these data are probably the dominant parameters adding uncertainties in the calculations of building energy performance [5]. In parallel, profiles allow a control system to guess the next steps in advance (e.g. [7], [17], [8]). Indeed, there is a high sensitivity in the user behavior or energy usage parameters, minor variations result in considerable differences. Unfortunately, often these inputs are not available and must be approximated, most of the times being not accurate [18].

As far as modeling the human behavior for the building optimization is concerned, absolute statements can hardly be established. It is often neither possible to determine exactly what the best model is nor to find the best method to obtain it. On the other hand, it is also difficult to abstract information from the discovered profiles or models. Moreover, it is even harder to assure that the profiles of a population will be suitable for another population (or give a measure of it). Also, only a few of today's buildings are monitored providing detailed information for a valid benchmarking. The scenario depicted above forces us to assume a certain level of uncertainty. We are convinced that the minimization of this uncertainty can only be achieved by intensive aggregation of real building data ending in comprehensive building usage information databases [6].

Clustering is the technique used to discover representative patterns within collected data. Seen from another point of view, it is the process of arranging samples into "sensible" clusters based on a pre-defined similarity (the similarity criterion is also part of the discussion regarding uncertainty). Due to the unsupervised nature of the clustering task, trying to find out the best clustering method for a certain scenario is not trivial. For instance, [22] have compared different clustering methods and show the difficulty to get absolute assessments

because of the lack of benchmarks, [16] discuss the dependence on the data set characteristics, or [21] claim for significant improvements in the algorithms.

The outlier presence is also an additional question that adds complexity to the discovery of behavioral profiles. In statistics, the presence of outliers indicates some kind of problem. Often it is related to a sample which does not fit into the model, or an error in a measurement. In our case, there is neither an absolute mathematical definition nor a ubiquitous method to state whether or not a measured sample is an outlier (there is no reason to assume always normal distributions). Thus, an outlier has a flexible performance that strongly depends on the scenario, the nature of the samples, the distribution, and what we expect from the classification.

[4] emphasizes the necessity of analyzing and determining the outlier origins, in order to know if the outliers are helping us to discover new knowledge ("good outliers") or they are just noise ("bad outliers"). We fully agree with his conclusion that it requires not only an understanding of the mathematical properties of data but also relevant knowledge in the domain context in which the outliers occur. But it is also possible that the classification results are needed to obtain some of this relevant knowledge in the domain context. It is not unusual that many variables take part in the performance of the profiles and part of them are usually not collected or cannot be easily collected or abstracted, or even are hidden or unknown.

Though, there is a broad experience using SOM for user and usage patterns. Therefore, we are convinced that SOM algorithms are also a good choice to perform usage profile discovery in the built environment. However, results from previous energy usage scenarios suggested the mathematical analysis of available building data and pointed out the outlier presence. This caused the necessity of refining SOM in order to be able to reject outliers. The final result is the development of XSOM.

3 eXclusive SOM (XSOM)

Self-organizing Maps, also called Kohonen networks [11], are a kind of unsupervised and competitive artificial neural network widely deployed for clustering. They allow a mapping from one (usually high-dimensional) space to another (usually low-dimensional) space. In previous applications, we have tested their flexibility, accuracy, local minima and variable density management capabilities in comparison with other methods [9]. XSOM [8][3] tends to overperform SOM solutions mainly when the deployment of clustering tools is intended for pattern discovery. Remote samples which are considered outliers for XSOM are classified by SOM without any distinction. Therefore, output patterns in the SOM case can be significantly different to the ones in the XSOM case due to the corruption introduced by outliers. In some cases (like in our case study), the outlier presence can seriously affect the shape of the representatives and even the SOM ability to identify good clusters.

[3] The XSOM algorithm has been previously depicted in [8] (as ESOM).

In order to achieve this filtering capability, the XSOM algorithm introduces a new parameter, called *tolerance*, that fixes the admitted level of appropriateness. In addition, XSOM also informs about the nearest cluster of each outlier. The main drawback of XSOM is the adjustment of the tolerance, it adds a new layer of complexity and discussion. Indeed, XSOM with a tolerance equal to infinite is just the normal SOM. Hence, SOM may be regarded as a specific case of XSOM.

4 Case Study: Water Consumption Profiling

4.1 Water Consumption Database

Our current research about energy usage profiles deploys the Leako System database. Leako is an enterprise from the Basque Country (Spain) specialized in central heating, Domestic Hot Water (DHW) and air conditioning installation, distribution, and metering. The Leako Database consists of hourly energy data obtained for seven years from more than 700 dwellings. The collected data comprise of heating (KWh), DHW (KWh), average indoor temperatures, and consumed amount of water. The latter was taken as database for this research.

The validity of profiles is based on the existence of trends and repetition patterns inside the data. The experience managing the database corroborates this statement but it is necessary to have a mathematical approach that supports it. For this reason, we started with an extensive analysis in order to validate the next assessments:

a) People keep habits.
b) Habits of some people are similar.

If every dwelling is analyzed independently and considered as a temporal series, stationary process criteria are not usually achieved and it does not allow to apply time series analysis and model estimation methods. Nevertheless, selecting a dwelling and analyzing the correlation between its days, after filtering absent days (no water consumption), the results very often conclude in a high number of correlated days (Pearson's correlation). Table 1 shows some collected data from random dwellings that corroborate the previous assessment.

Table 1. Correlation (ρ) between days for three dwellings selected at random.

	Dwelling 1	Dwelling 2	Dwelling 3
Total days	2641	2634	1494
$af0$ (after filtering 0-days)	2384	1527	1274
$\rho_{0.6}(500)_{af0}{}^{*}$	1001	199	526
$\rho_{0.9}(100)_{af0}{}^{**}$	567	142	179

**: days that have a $\rho \geq 0.6$ with more than 500 other days.*
***: days that have a $\rho \geq 0.9$ with more than 100 other days.*

If data concerning each dwelling are condensed or summarized through statistic procedures (or clustering tools), it is also possible to study the correlation between different dwellings in order to know if there exist similar habits between people. Some results are shown in Table 2.

Table 2. Correlation (ρ) between dwellings

Total dwellings	685
$af0$ (after filtering 0-days)	668
$\rho_{0.6}(50)_{af0}{}^{*}$	426
$\rho_{0.9}(5)_{af0}{}^{**}$	74

**: dwellings that have a $\rho \geq 0.6$ with more than 50 other ones.*
***: dwellings that have a $\rho \geq 0.9$ with more than 5 other ones.*

For the model obtaining and the clustering tool preparation, the monitored raw data of the database are subjected to some transformations:

1. 0-days and incomplete data are removed.
2. Following the Spanish Technical Code for Buildings (CTE) characteristics [3], each dwelling is represented by a frame of 63 (7x3x3) cells. This transformation allows to map data to seven days a week, grouped in three periods of a day, classified by three seasons. Each cell shows the consumed liters per hour in a period of 8 hours.
3. Each dwelling consumption is related to a hypothetical user. This estimation is obtained from the Spanish Ministry of Health and Consumer Affairs.
4. Regardless of the fact that the data distribution is much narrower than in a normal distribution case, a second normalization – using mean and standard deviation – is executed [11].

4.2 Experiments and Parametrization of Clustering Methods

In order to evaluate XSOM for the usage pattern discovery in a building profiling case with outliers, different tests and comparisons have been undertaken.

XSOM is tested executing a sweep of tolerances and assessing the different performances. Later on, the performances are compared with SOM and K-means clustering methods. Except for the tolerance variation, the rest of the parameters have been equivalently chosen for SOM and XSOM. The application demands big spherical and globular clusters with high representativeness (marginal dense clusters are not important). The initial number of clusters has been fixed to 5 according to the maximum desired for the application and taking into account that SOM method does not present noticeable variations with a number of initial clusters between 5 and 10. The similarity function is based on Euclidean distance.

With regard to K-means method (using CLUTO tools [10]), the best performance is reached applying direct K-means methodology, with an initial number

of 5 clusters, using Euclidean distances and the I_2 criterion for the optimization function. Different parametrizations and other approaches, like Repeated Bisection or Graph arrangements, have also been tested with worse results.

For the evaluation (i.e. the validity of the clustering method), the following outputs in each performance have been studied: *a)* Number of significant patterns (Ps). *b)* Number of outliers (nO). *c)* Form of patterns. *d)* Number of samples embraced in each cluster (nP_j, where j is the pattern identifier in the corresponding experiment). *e)* Distances between representatives. *f)* Distances and statistical data between representatives and their embraced samples.

4.3 Results

Table 3 shows a number of results obtained by different tolerance factors. As long as tolerance decreases, the number of outliers grows (Fig. 1). In parallel, the number of significant patterns rises. This is due to the disappearance of the outlier distortion, that appears when they are classified and accepted inside clusters. Outliers move the gravity center of the group and, thus, decisive differences between close elements are ignored. While SOM only detects one pattern, XSOM configurations detect more significant groups. In addition, the appearance of new clusters is also due to the fact that tolerance adjustment redefines the meaning of "outlier". Tolerance fixes how far samples can remain from the cluster center, so outliers are not only errors or samples that distort normal distributions. Summarizing, XSOM allows an outlier to be part of a new group of non-clustered members.

Table 3. Results in the tolerance sweep test

tol_m	Ps	nO	$\%O$	nP_1	nP_2	nP_3	nP_4	nP_5
∞	1	0	0.0%	659	13	4	6	2
158.70	1	23	3.4%	648	4	4	3	2
79.37	2	31	4.5%	538	110	4	1	0
39.68	2	43	6.3%	519	122	0	0	0
15.87	2	60	8.8%	404	220	0	0	0
7.94	3	72	10.5%	352	222	38	0	0
3.97	3	93	13.6%	376	191	24	0	0
1.59	3	201	29.4%	223	192	68	0	0
0.79	3	307	44.9%	184	97	95	0	0
0.32	4	541	79.1%	64	30	26	23	1

Paying attention to the distance values between the representative pattern and the samples in SOM, the existence of outliers can be confirmed (according to the common definition). Distances do not follow a normal distribution but their relationship is close to a logistic distribution that resembles normal distribution in shape but has a higher kurtosis. Most of the variance is due to odd extreme deviations [1]. A logistic distribution can be expressed as follows:

$$f(x, \sigma, \mu) = \frac{1}{2} + \frac{1}{2} tanh\left(\frac{x-\mu}{2\sigma}\right) \tag{1}$$

where μ stands for the mean and σ is the standard deviation. The usual definition of the (excess) kurtosis (γ_2) is shown in the next equation:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \tag{2}$$

where μ_4 is the fourth moment about the mean. Whereas in a normal distribution the excess kurtosis equals 0, in a logistical distribution it equals 1.2.

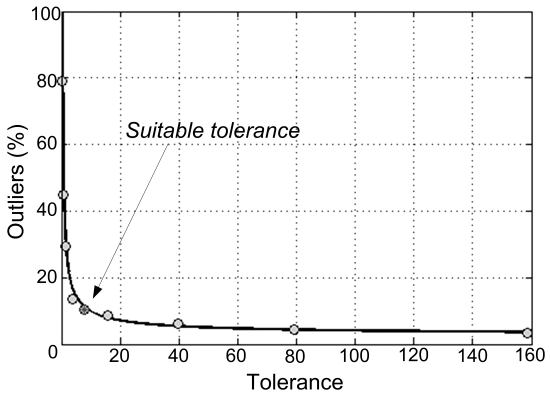

Fig. 1. Outliers vs tolerance

It is desirable to reach a good tolerance value that rejects a suitable number of outliers. Outliers-tolerance relationship fits an inversely proportional function (Fig. 1). According to this relationship, the compromise is reached when the increment of outliers and the increment of tolerance are balanced ($\Delta out = \Delta tol$). The tolerance that matches the previous conditions is close to 0.0158 (fifth experiment, Table 3) resulting in 8.8% of outliers in the whole population. In that case, XSOM identifies two patterns. The criterion applied to establish the selected tolerance can be widely discussed because it has been stated without a previous outlier definition and based on a commitment between the tolerance-outlier evolution. In any case, the tolerance sweep, in its medium values, always shows two main patterns that do not differ too much in shape and members.

The most remarkable question concerns the differences between SOM and XSOM comparison. In Fig. 2, SOM is compared against the most suitable XSOM performance (with a tolerance equal to 0.0158). While SOM considers a great group (96.3% of input samples) the XSOM classification delivers two groups (with 59.1% and 32.2% elements, respectively). Thus SOM ignores a well-defined group that represents 32% of the population in XSOM classification. Instead of

that, it absorbs these elements into the group represented by the big pattern. As it can be noticed in Fig. 2, the outlines of most significant patterns in SOM and XSOM experiments are similar, but SOM has higher values in the whole range.

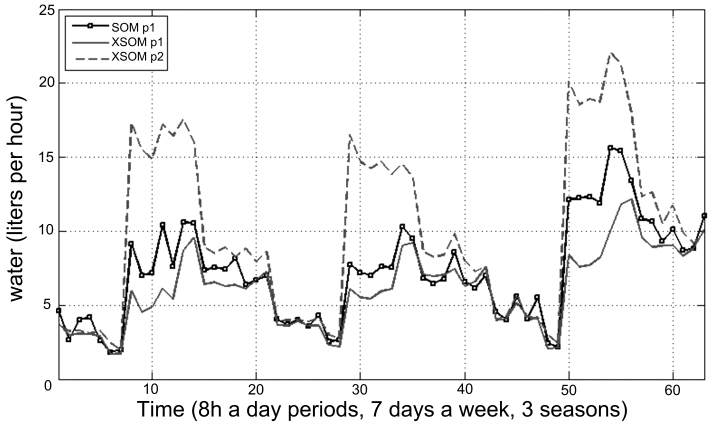

Fig. 2. SOM water pattern

In the K-means case, the two main patterns embrace 42.3% and 24.1% of the inputs, respectively. The different performances are assessed based on the average similarities and standard deviations between the obtained patterns and the patterns with their respective clustered samples. In any case, resulting representatives show distorted and even incoherent shapes. We conclude that approaches based on K-means methodology are poor to deal with the current scenario.

Without benchmarks or a a validation technique, it is difficult to establish a best method, but it is possible to submit the results to statistical analysis in order to know how close the input samples are to their representative patterns (internal cluster validity). Table 4 shows the evaluation. It discloses that elements classified in XSOM are very much closer to their own patterns than elements in SOM or K-means classifications. Therefore, the two patterns obtained in XSOM case are more representative than the others.

So far, Leako's technical experts confirm the existence of two trends in users. Besides water consumption, this also concerns heating and cooling energy consumption (backed up by the billing). These trends have not yet been carefully studied – this could also be due to geographic matters, building or family sizes, orientation, different systems, or other reasons.

5 Conclusions and Discussion

This paper studies the convenience of XSOM clustering for pattern discovery in a behavioural profiling case where outliers are existent. While in such a case

Table 4. Internal cluster validity in the three cases.

		SOM	XSOM	K-means
P1	samples	96.3%	59.1%	42.3%
	distance (mean)	0.61	0.06	0.63
	distance (σ)	5.59	0.07	0.87
P2	samples	–	32.2%	24.1%
	distance (mean)	–	0.12	2.39
	distance (σ)	–	0.17	0.44

SOM may fuse significant clusters and distort the representatives, XSOM is able to filter erratic data, obtain best representatives and identify members that must not be classified into any group. In other respects, XSOM also allows to apply other SOM enhancements and evolutions (e.g. [2]).

The main drawback of XSOM is the necessity to adjust a tolerance parameter. In the discussed use case, it has been deduced assuming that when the sensitivity regarding tolerance starts increasing quickly the clustering methodology is beginning to reject non-erratic samples.

The tolerance parameter introduces the concept of *focusing* in clustering. Perhaps it is difficult, or even impossible, to establish the right tolerance value in certain cases . In other words, an ambiguous scenario can admit different clustering solutions. Therefore, XSOM would improve SOM because it allows focusing or diverse granularity interpretations. In addition, tolerance sweeps imply a new hierarchical clustering approach that does not impose clustering shapes to the data as much as other agglomerative methods do [19]. Nevertheless, setting the right degree of tolerance for a given scenario is not a trivial issue which needs to be thoroughly analyzed in future work.

Despite the uncertainties and difficulties, it is worth applying clustering methods to discover behavior profiles for building energy performance and smarter control [17]. For instance, the water profiles obtained with XSOM have already been used to reach more realistic energy simulations in Spanish buildings (using Calener) and, combined with electricity usage profiles, for modeling occupancy in control application studies. Next planned activities include to deploy them to model DHW management systems and to optimize a predictive production.

Acknowledgements The work presented in this paper was funded by the HdZ+ programme of the Austrian Research Promotion Agency FFG under the project 822170.

References

1. Balakrishnan, N.: Handbook of the Logistic Distribution. Marcel Dekker, New York (1992)
2. Berglund, E., Sitte, J.: The parameterless self-organizing map algorithm. Neural Networks, IEEE Transactions on 17(2), 305–316 (2006)

3. Cantos, S., Iglesias, F., Vidal, J.: Comparison of standard and case-based user profiles in building's energy performance simulation. In: Building Simulation 2009. Eleventh International IBPSA Conference. pp. 584–590 (2009)
4. Cheng, J.G.: Outlier management in intelligent data analysis. Ph.D. thesis, University of London (2000)
5. Corrado, V., Mechri, H.E.: Effect of data uncertainty on energy performance assessment of buildings. In: Climamed 2007 Proceedings. pp. 737–758 (2007)
6. Crosbie, T., Dawood, N., Dean, J.: Energy profiling in the life-cycle assessment of buildings. Management of Environmental Quality: An International Journal 21, 20–31 (2010)
7. Heierman, E.O., I., Cook, D.: Improving home automation by discovering regularly occurring device usage patterns. In: Data Mining, Third IEEE International Conference on. pp. 537–540 (2003)
8. Iglesias Vázquez, F., Kastner, W.: Usage profiles for sustainable buildings. In: Emerging Technologies and Factory Automation, 2010 IEEE Conference on. pp. 1–8 (2010)
9. Iglesias Vázquez, F., Kastner, W.: Clustering methods for occupancy prediction in smart home control. In: Industrial Electronics, 2011 IEEE International Symposium on. p. unpublished (2011)
10. Karypis, G.: CLUTO: A Clustering Toolkit. University of Minnesota, Dept. of Computer Science, Minneapolis, MN (2003), release 2.1.1
11. Kohonen, T.: The self-organizing map. Proceedings of the IEEE 78(9), 1464–1480 (Sep 1990)
12. Lingras, P., Hogo, M., Snorek, M.: Interval set clustering of web users using modified kohonen self-organizing maps based on the properties of rough sets. Web Intelligence and Agent Systems 2, 217–225 (August 2004)
13. Mahdavi, A., Pröglhöf, C.: User behaviour and energy performance in buildings. In: IEWT 2009, Int. Energy Economics Workshop TUV. pp. 1–13 (2009)
14. Mills, E.: Inter-comparison of north american residential energy analysis tools. Energy and Buildings 36(9), 865–880 (2004)
15. Mingoti, S.A., Lima, J.O.: Comparing som neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. European journal of operational research 174(3), 1742–1759 (2006)
16. Qian, W., Zhou, A.: Analyzing popular clustering algorithms from different viewpoints. Journal of software 13(8), 1382–1394 (2002)
17. Reinisch, C., Kofler, M.J., Iglesias, F., Kastner, W.: ThinkHome: Energy efficiency in future smart homes. EURASIP Journal on Embedded Systems 2011, 1–18 (2011)
18. Sabaté, J., Peters, C.: 50% CO2-reduction in Mediterranean social housing through detailed life-cycle analysis. In: Climamed 2007 Proceedings. pp. 927–959 (2007)
19. Theodoridis, S., Koutroumbas, K.: Pattern Recognition: Fourth Edition. Academic Press (ELSEVIER) (2003)
20. Verdu, S., Garcia, M., Franco, F., Encinas, N., Marin, A., Molina, A., Lazaro, E.: Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters. In: Power Systems Conference and Exposition, 2004 IEEE Conference on. vol. 2, pp. 899–906 (2004)
21. Wu, J., Hassan, A.E., Holt, R.C.: Comparison of clustering algorithms in the context of software evolution. vol. 0, pp. 525–535. IEEE Computer Society, Los Alamitos, CA, USA (2005)
22. Zheng, X., Cai, Z., Li, Q.: An experimental comparison of three kinds of clustering algorithms. In: Neural Networks and Brain, 2005. ICNN B '05. International Conference on. vol. 2, pp. 767–771 (2005)