Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 122 (2017) 229–234

Energy Procedia

www.elsevier.com/locate/procedia

CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale, CISBAT 2017 6-8 September 2017, Lausanne, Switzerland

Smart Cities (Urban Simulation, Big Data)

Unsupervised load shape clustering for urban building performance assessment

Jimeno A. Fonseca[a,b*], Clayton Miller[a,c], Arno Schlueter[a,b]

[a]*Future Cities Laboratory (FCL), Singapore-ETH Centre (SEC), 1 Create Way, Singapore 13892, Singapore*
[b]*Architecture and Building Systems (A/S), ETH Zurich, John-von-Neumann-Weg 9, Zurich 8093, Switzerland*
[c]*Building and Urban Data Science (BUDS) Lab, National University of Singapore (NUS), 4 Architecture Drive, 117566 Singapore*

Abstract

This paper presents a method to automatically cluster typical days of energy consumption in one or several buildings. The method is based on an optimized version of the Symbolic Aggregate approXimation (SAX) method. SAX is a data mining technique for clustering time series with recent applications in building fault detection and building performance assessment. The number of clusters and accuracy of SAX highly depends on two highly sensitive input variables, i.e., the word size and the alphabet size. We propose the use of the genetic algorithm NSGA-II to optimize the number of words and alphabet size of SAX subjected to three fitness objectives, i.e., maximize data accuracy and compression and minimize complexity. In addition, we propose the use of MAVT as selection method of the optimal solution. The methodology is applied to measured energy consumption data of three representative buildings on a university campus in Singapore. Potential future uses of the approach include advanced studies in fault detection and calibration of urban building performance models.

Peer-review under responsibility of the scientific committee of the CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale

Keywords: Building performance; Data mining; Daily Profile Extraction;

* Corresponding author.
E-mail address: fonseca@arch.ethz.ch

1876-6102
Peer-review under responsibility of the scientific committee of the CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale
10.1016/j.egypro.2017.07.350

1 Introduction

Grouping similar behavior in time-series data is a well-established process to characterize the way buildings use energy. The practice of daily load profile clustering is a specific technique to group building performance data. This procedure begins with the collection and cleaning of raw, sub-hourly sensor data from electrical meters or utility smart meters. Sub-hourly building performance data are chunked into 24-hour sub-sections. These diurnal patterns are then clustered by comparing the measurements for all samples at each hour or sub-hourly mark of the day and using distance metrics to group profiles that are roughly similar. Several unsupervised clustering techniques have been tested in their application to performance measurement data from buildings. A seminal review of clustering techniques load pattern grouping provides an overview of phases of analysis, clustering techniques, and an in-depth discussion of quantified validation and comparison of several tested techniques [1]. Other studies focused on the comparison of clustering algorithms and distance metrics as applied specifically to load profiles are cited in [2], [3]. Research outlining the application of load profiling in non-residential buildings includes its use for improving load forecasting [4] and utility grid analytics [5]. One key approach of [6] defined as the *DayFilter* process, was recently developed as a specific diurnal load profile clustering and filtering process for building energy consumption. The method utilizes the Symbolic Aggregate approXimation (SAX) clustering technique. In contrast to other clustering techniques, SAX has the advantage of speed and ease of use on large time-series data sets [7] [8]. However, SAX requires several inputs such as the size of the generated word (sub-sequence window size) and alphabet size (magnitude breakpoints) as input variables and manual selection of these values is onerous when applying the process to large groups of buildings.

This research built upon previous work by refining the use of measured load shape profiling in the domain of urban-scale building performance simulation. The approach consists of automating the SAX clustering technique with the use of multi-objective optimization for selection of the word and alphabet size in SAX. For demonstration purposes, the method is applied to hourly energy consumption of three typical buildings of the Nanyang Technological University (NTU) Campus of Singapore.

1.1 Data collection and processing

Sub-hourly metered data between 28.08.2014 – 07.07.2016 was gathered from the BMS of each building and cleaned with a time-of-week and temperature (TWOT) model *[9]*. TWOT served to remove outliers and fill in gaps in the data. Fig. *1* illustrates typical raw data for three of the most common use types in the area of study.

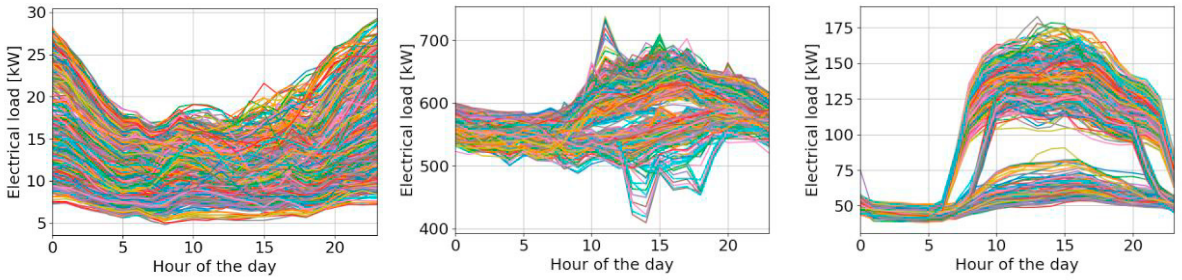

Fig. 1: Raw daily load profiles from three typical building use types. A. dorm, b. laboratory, c. office

1.2 Clustering

The SAX process converts time-series data into a set of strings that can then be grouped together for the purposes of clustering, motif and discord analysis, and finding anomalous behavior. Fig. 2 illustrates the SAX process as implemented on two days of data. Within the context of daily load profile, each day is divided into 24-hour periods starting at midnight. Each day is then further divided into segments based on the input parameter *w,* or window size. This parameter dictates how many sub-sequences are contained within a day period, and thus, how much data is averaged to create each letter of a string and how many letters long the string will be. The magnitude of the measured

variable is also divided into a number of regions according to the parameter a, or alphabet size. Each subdivided region is assigned a letter of the alphabet and this letter is used if the mean value of the data for each subsequence window falls in that region. The two days in Fig. 2 follow this process with the parameter settings of $w = 4$, and $a = 3$. The strings *acba* and *abba* were created for these two days. The clustering process simply groups all daily profiles of a certain string to create bins of samples that are somewhat similar. SAX can modulate between different levels of aggregation and detail by choosing different input parameter settings. The larger *a* or *w* are, the lower the number of clusters are.

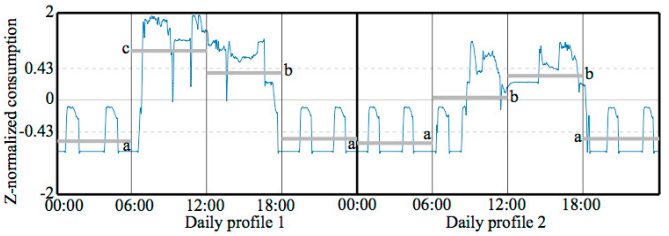

Fig. 2: Example of the SAX conversion of two days of building performance data. The each day is divided into four segments of six hours (w=4) and the magnitude is divided into three parts (a=3) (Used with permission from [6])

1.3 Multi-objective optimization

Defining the optimal number of clusters in time-series data is a multi-objective optimization problem. In terms of the SAX method, the inputs *w* and *a* should be selected to minimize the number of different strings (clusters) while maximizing the similitude and complexity of data represented in each cluster. Acosta et al., [10] conclude that these conflicting objectives can be represented by three main performance metrics: accuracy, complexity, and compression. Accuracy (A) (Eq.1) describes how reliable a variable is to predict the value of other variables. It is computed through a normalized entropy calculation of n' strings. Where P_i is the probability of occurrence of string i, and $n = 365$ are the number of days included in the original time series.

$$\mathrm{A} = -\frac{\sum_{i=1}^{n'} P_i \cdot log_2(P_i)}{log_2(n)} \tag{1}$$

Complexity (B) limits the generation of different strings (clusters) by penalizing the discretization scheme. It is computed according to Eq. 2. In contrast to Acosta et al., [10], we use the inverse of complexity to convert the objective function of Eq. 4 in an entire maximization problem. A complexity close to one is equivalent to a low number of clusters.

$$\mathrm{B} = 1 - n'/n \tag{2}$$

Compression (Γ) accounts for the capacity of the clustering technique to reduce the dimensionality of the $k = 24\,h$ period selected to represent a day. It is a measure of the compression rate of a particular discretization scheme $\{w, a\}$. Values of compression closer to one are equivalent to higher compression rates or a lower number of w in every string.

$$\Gamma = 1 - w/k \tag{3}$$

The genetic algorithm type NSGA-II is selected to find the global optima of the multi-objective optimization problem. The genetic algorithm is one of the most efficient and widely validated [11]. We tested the algorithm for 100 individuals and 500 generations finding convergence (i.e., a stable diversity = 0.982) after 100 generations.

$$O.F = \max(\mathrm{A}, \mathrm{B}, \Gamma) \tag{4}$$

1.4 Selection

In order to select the most optimal solution of the Pareto-Frontier, we propose a Multi-Criteria Decision Analysis based on the multi-attribute value theory (MAVT). MAVT computes a global value (V) out of the partial values u of the criteria under analysis (Eq.4) [12]. The equation is applied over every individual of the Pareto-front and the individual with the maximum global value is selected. For demonstration purposes we applied the arbitrary weights in the scale of 10 – 100 of $\alpha = 100, \beta = 100, \gamma = 70$ for $A, B, and\ \Gamma$ respectively. The calibration of these weights remains part of future work (see [10]). Figure 3 illustrates the results for three typical building use types in the area of study.

$$V = \sum_{i=1}^{3} W_i \cdot u_i \ \ where: u_i \in \{A, B, \Gamma\}, \qquad W_i \in \{\alpha, \beta, \gamma\} \tag{5}$$

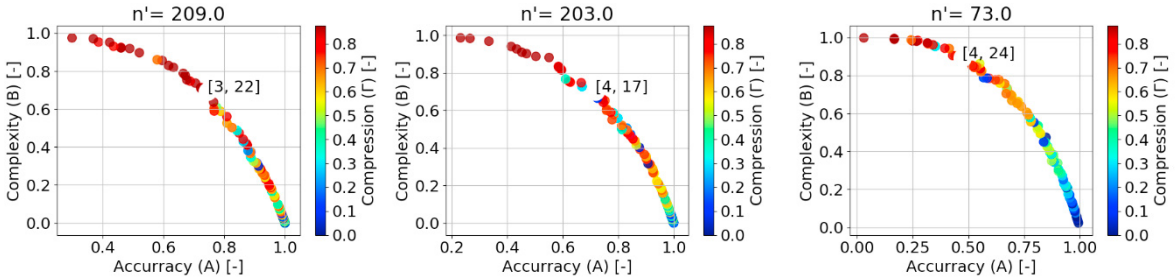

Figure 3: Pareto-frontier and selected individual for typical building use types. A. dorm, b. laboratory, c. office

2 Results

Figure 4 presents the daily load shape of the three building types of the study. The figure compares the results of the present study (optimization case) with those of a traditional application of SAX in building performance analysis (default case) where $w = 4$, and $a = 3$ (see section 1.2). For the building uses of interest the optimization case requires close in the range of seven to 13 times more clusters than the default case. Despite of this, data accuracy increases between 30% and 40% for the optimization case. Table 1 presents a summary of performance indicators of the default case and optimization case.

Data	Default case							Optimization case					
Building Use Type	n	w, a	n'	A	B	Γ	V	w, a	n'	A	B	Γ	V
Dormitory	680	3, 4	20	0.37	0.97	0.88	70.7	3, 22	209	0.74	0.69	0.88	75.8
Laboratory	680	3, 4	12	0.26	0.98	0.88	66.7	4, 17	203	0.71	0.70	0.83	74.4
Office	680	3, 4	4	0.05	0.99	0.88	66.7	4, 24	73	0.47	0.89	0.83	74.8

Table 1 Comparison in gains of indicators of default case vs. optimization case.

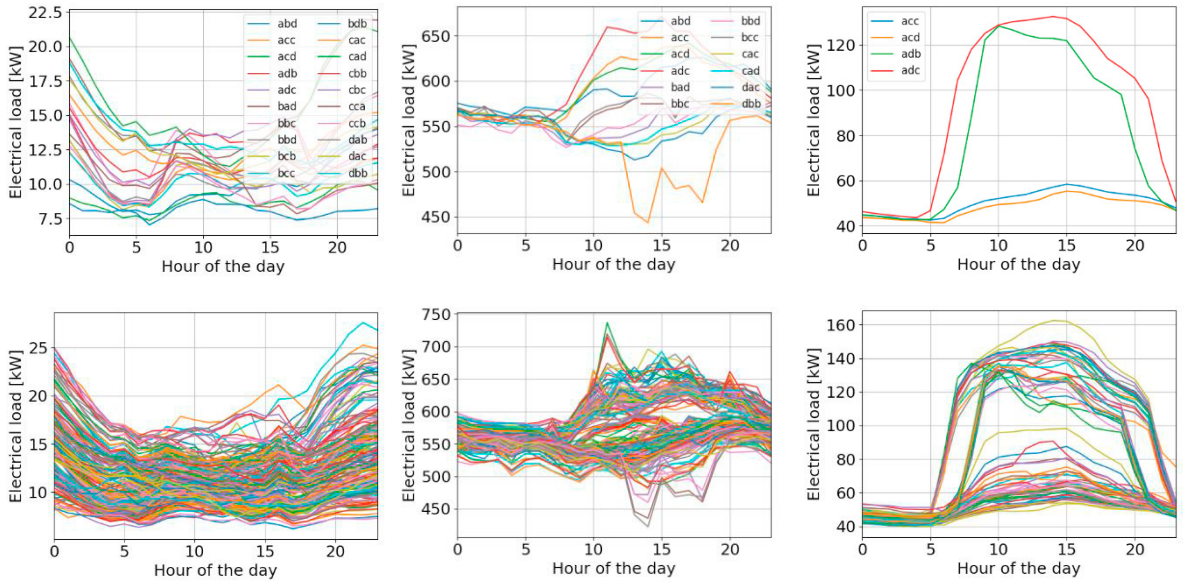

Figure 4: Clustered mean daily load profile for dormitory, laboratory, and office. Top: Default case: Bottom: Optimize case. From left to right: Dormitory, Laboratory, and Office.

3 Conclusions

This research built upon previous work by refining the use of measured load shape profiling in the domain of urban-scale building performance assessment. Specifically, this work builds upon the use of SAX as a technique for quickly clustering daily load profiles. Several key novelties come out of this work. The first is the automated parameter selection of load profiles for different building use types. The second is a new unsupervised method based on the combination of SAX, NSGA-II, and MAVT. The method is coded in an open source language and made available as a module for the City Energy Analyst tool[13].

For the case study, the results depict an increase in 35% - 45% and 10 - 30% for data accuracy and complexity in comparison to the default case (w =3, a=4) of SAX in building performance simulation. Despite this, the optimal selection of clusters might be lacking a similitude index to constrain groups of similar clusters.

This research is important for the purpose of characterizing diurnal building performance behavior. This information is useful for several applications related to whole building and district-scale building simulation, optimization of the design of large-scale heating and cooling systems, and detection of anomalous behavior that could be attributed to performance improvement opportunities. Typical daily performance behavior is used to create utilization and diversity schedules in simulation programs such as the City Energy Analyst (CEA) and EnergyPlus. The automated creation of these profiles enables the scaled analysis of several buildings of buildings in large urban-scale groups. The availability of data at these scales is increasing rapidly due to new sources such as campus energy management systems and smart meters at the grid-scale. Anomaly detection at these scales can also be greatly enhanced through automated load profile generation. The measured performance of a specific building can be compared to its peers not only in terms of magnitude but also in its typical patterns of use. This type of analysis detects the misclassification of a building's primary use; it can help understand, for example, whether a building that is labeled as an office is “acting” more like a laboratory in its use patterns. If so, it might be the case that the building has been renovated or has equipment more consistent with a more intensive use type. This information is useful when determining which buildings are its peers and what standard that building should be compared against.

Further research in this direction lies in the use of the generated profile patterns for the calibration criteria for MAVT, and the study of similitude indexes.

Acknowledgements

This research has been financed by the National Research Foundation (NRF) of Singapore under the Future Cities Laboratory (FCL). The authors would like to thank the EcoCampus team and Energy Research Institute at NTU (ERI@N) at the Nanyang Technological University (NTU) of Singapore for support on data collection.

References

[1] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, May 2012.
[2] F. Iglesias and W. Kastner, "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns," *Energies*, vol. 6, no. 2, pp. 579–597, Feb. 2013.
[3] S. Ramos, J. Duarte, J. Soares, Z. Vale, and F. J. Duarte, "Typical load profiles in the smart grid context—A clustering methods comparison," in *Proceedings of the Power and Energy Society General Meeting, 2012 IEEE*, San Diego, CA, USA, 2012, pp. 1–8.
[4] A. Shahzadeh, A. Khosravi, and S. Nahavandi, "Improving load forecast accuracy by clustering consumers using smart meter data," in *Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, 2015, pp. 1–7.
[5] R. Green, I. Staffell, and N. Vasilakos, "Divide and Conquer K-Means Clustering of Demand Data Allows Rapid and Accurate Simulations of the British Electricity System," *IEEE Trans. Eng. Manag.*, vol. 61, no. 2, pp. 251–260, May 2014.
[6] C. Miller, Z. Nagy, and A. Schlueter, "Automated daily pattern filtering of measured building performance data," *Autom. Constr.*, vol. 49, Part A, pp. 1–17, Jan. 2015.
[7] J. Lin, E. J. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," *Data Min. Knowl. Discov.*, vol. 15, no. 2, pp. 107–144, 2007.
[8] C. Miller and A. Schlueter, "Forensically discovering simulation feedback knowledge from a campus energy information system," in *Proceedings of the 2015 Symposium on Simulation for Architecture and Urban Design (SimAUD 2015)*, Washington DC, USA, 2015, pp. 33–40.
[9] J. L. Mathieu, P. N. Price, S. Kiliccote, and M. A. Piette, "Quantifying Changes in Building Electricity Use, With Application to Demand Response," *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 507–518, Sep. 2011.
[10] H.-G. Acosta-Mesa, F. Rechy-Ramírez, E. Mezura-Montes, N. Cruz-Ramírez, and R. Hernández Jiménez, "Application of time series discretization using evolutionary programming for classification of precancerous cervical lesions," *J. Biomed. Inform.*, vol. 49, pp. 73–83, Jun. 2014.
[11] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.
[12] A. Azapagic and S. Perdan, "An integrated sustainability decision-support framework Part II: Problem analysis," *Int. J. Sustain. Dev. World Ecol.*, vol. 12, no. 2, pp. 112–131, Jun. 2005.
[13] J. A. Fonseca *et al.*, "City Energy Analyst 2.2," 2017. https://doi.org/10.5281/zenodo.556165