УДК 004.9

Олег Сінькевич[1], Любомир Монастирський[1], д. ф.-м. н., проф., Богдан Соколовський[1], к. ф.-м. н., доц., Зіновій Матчишин[2]

[1]Львівський Національний Університет ім. І. Франка, Україна,
[2]Altran/Lohika, Україна

КЛАСТЕРНИЙ АНАЛІЗ ЕНЕРГЕТИЧНИХ ЧАСОВИХ РЯДІВ РОЗУМНОГО БУДИНКУ

З настанням та швидким розвитком четвертої індустріальної революції та інтелектуалізацією функціонування різних об'єктів, зокрема, розумних будинків, виникає потреба у розробці та дослідженні різноманітних алгоритмів оптимізації енерговитрат. В даній роботі для виявлення та оцінки закономірностей енергоспоживання здійснено аналіз часових рядів споживання газу для розумних будинків. З використанням алгоритмів машинного навчання проведена кластеризація річних даних газу та виявлені добові патерни відносно різних сезонів протягом року. На основі отриманих патернів проаналізовано поведінкові залежності споживання енергії, яка використовується для обігріву та запропоновано підхід до дисагрегації даних споживання газу.

Ключові слова: Кластеризація, розумний будинок, часові ряди, машинне навчання.

Oleh Sinkevych, Liubomyr Monastyrskyi, Bohdan Sokolovskyi, Zenyk Matchyshyn

CLUSTER ANALYSIS OF SMART HOME ENERGY TIME SERIES

With the rapid development of Industry 4.0 and the intellectualization of functioning the various objects, in particular, smart homes, there is a need for the research and development of various algorithms for energy optimization. In this paper, for detecting and evaluating the patterns in energy data, the time series of smart home gas consumption have been analyzed. Using machine learning algorithms, clustering of annual data has been performed and daily patterns have been detected for different seasons during the year. On the basis of the obtained patterns, behavioral dependencies of energy consumption used for heating have been studied an approach to gas time series disaggregation has been proposed.

Keywords: Clustering, smart home, time series, machine learning.

1. Introduction. The rise of the different intelligent automation systems is inextricably linked to the emergence of the modern mathematical methods like statistical and machine learning algorithms [1]. The Internet of things (IoT) technology, which is one of the main research subjects in commercial and academic fields, requires incessant improvements due to the quick growth of computational resources and appearance of high-powered single-board computers [2]. Smart home as a part of IoT continues to be an object of interest among other technologies and is also considered as the element of wider smart city solutions. While speaking about smart home, we almost always bear in mind an energy saving problem. It covers the problem of effective energy management system [3], the understanding of resource consumption as well as the forecasting and activity pattern recognition.

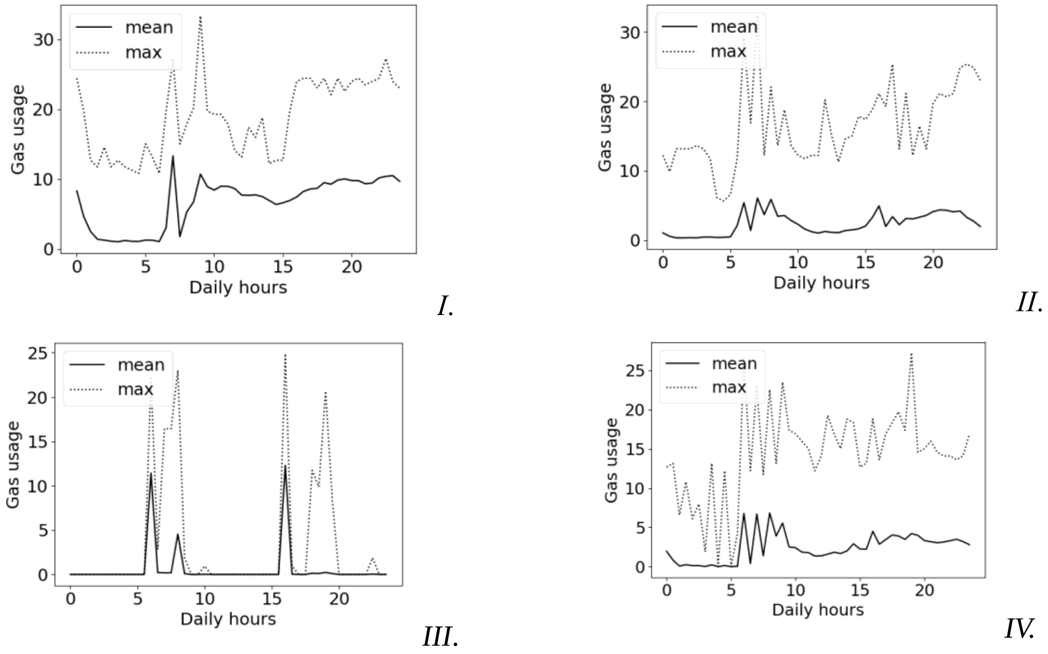
In this paper, we are studying the second problem mentioned in previous paragraph – the understanding of energy consumption in a domestic household. Here under energy consumption we mean the total gas usage by residents in a single-family detached home. The aims of the conducted research are to: 1) calculate different clusters for daily gas consumption based on seasonality; 2) detect clusters in gas time series and 3) propose an approach for gas data disaggregation. The latter means splitting total gas data into components used for heating and non-heating (cooking, having a shower, etc.) purposes.

In section 2, the data processing and visualization are provided; a clustering method and its results are shown and discussed in section 3; an approach of time-windowing to detect signatures and events for the disaggregation problem is given in section 4.

2. Data preprocessing and exploration. In order to investigate our algorithms, we use the gas consumption data and temperature distributions (time series) from the open-access dataset, which was collected via smart meters under UK Refit Smart Home project [3].

Let's define the incoming set of gas time series as $\mathbf{G}_e = \left[x_e^{(1)}, x_e^{(2)}, \ldots, x_e^{(n)} \right]$ and outdoor temperature as $\mathbf{I}_t = \left[x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)} \right]$, where $x_e^{(1)}$ is the instant power value [W] converted from cubic meters [3] and $x_i^{(1)}$ is the corresponding value of outdoor temperature [${}^{o}C$]. To decide which type of data normalization should be applied, we have tested the gas time series for normality by the Shapiro-Wilk method [4]. It did not confirm null-hypothesis; therefore the minimum-maximum normalization algorithm has been applied: $x_e^{(j)} := \left(x_e^{(j)} - \min\left(\mathbf{G}_e\right)\right) \Big/ \left(\max\left(\mathbf{G}_e\right) - \min\left(\mathbf{G}_e\right)\right)$.

Next step is splitting the data into four seasonal time series, i.e., $\mathbf{G}_e^{I}$, $\mathbf{G}_e^{II}$, $\mathbf{G}_e^{III}$ and $\mathbf{G}_e^{IV}$, where I, II, III, IV are the index notations corresponding to winter, spring, summer and autumn periods. After that, each of the seasonal time series has been represented by daily sets resampled upon 30 minutes values per day (totally, 48 points). The idea of such a splitting is to identify the daily heating patterns observed during hot, cold and transitional seasons. Figure 1 demonstrates the typical maximum and average daily gas consumption by single family household in the moderate climate zone before normalization.



I. *II.* *III.* *IV.*

Fig. 1. Mean and maximum daily gas usage (in arbitrary units) in each season (winter, spring, summer, autumn)

As it is seen from Figure 1, the winter gas data (*I*) has pronounced peak in the morning hours, linear pattern during the day and slump during the night. It can be explained by the presence of the domestic morning activities by residents. We define this peak period as time interval $\left|\left[\mathbf{G}_e^{I}\right]_{ma}\right| \approx 1.5-2$ of length 1.5-2 hours. The spring and autumn time series (*II*, *IV*) share similar morning patterns with upturn peaks (afternoon activities) – $\left|\left[\mathbf{G}_e^{\{II,IV\}}\right]_{aa}\right| \approx 1.5-2$ in the second part of the day after 3 p.m. The summer data (*III*) consist of zero consumption except morning and afternoon activities found in *I, II* and *IV* time series: $\left|\left[\mathbf{G}_e^{III}\right]_{ma}\right| \approx 1.5-2$ and $\left|\left[\mathbf{G}_e^{III}\right]_{aa}\right| \approx 1.5-2$ correspondingly.

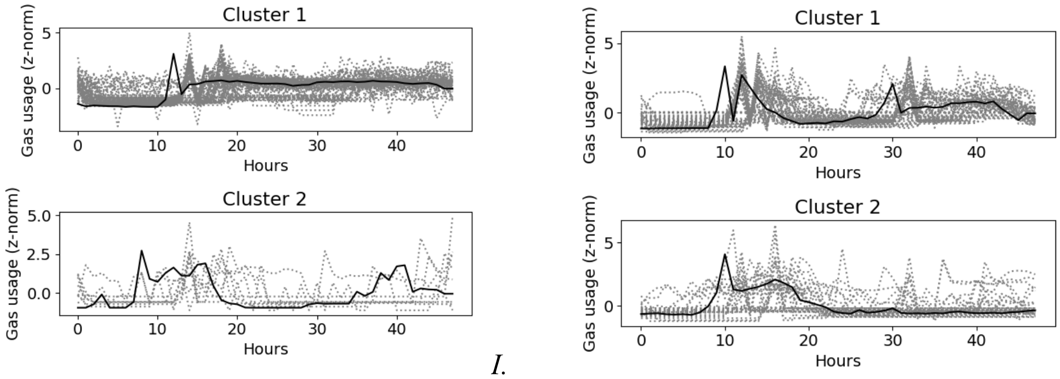
The extracted information from the visual exploration can be set as a basis for subsequent cluster analysis to find dissimilarities between the time series during each considered season.

3. Cluster analysis. The aim of clustering as a subfield of unsupervised learning is to split and group data items with respect to their similarity measured in variety of metrics. In our research, we have chosen a novel k-Shape clustering algorithm with the Euclidian metric, which shows very good results and has proven its effectiveness during testing:

$$\boldsymbol{\mu}_k^* = \underset{\boldsymbol{\mu}_k}{\arg\max}\left(\frac{\boldsymbol{\mu}_k^{\mathrm{T}} \cdot \mathbf{M} \cdot \boldsymbol{\mu}_k}{\boldsymbol{\mu}_k^{\mathrm{T}} \cdot \boldsymbol{\mu}_k}\right), \tag{1}$$

where $\boldsymbol{\mu}_k$ is the squared similarities between the time-series sequences, $\mathbf{M} = \mathbf{Q}^{\mathrm{T}} \cdot \mathbf{S} \cdot \mathbf{Q}$, $\mathbf{Q} = \mathbf{I} - \mathbf{O}/m$, $\mathbf{I}$ is the identity matrix, $\mathbf{O}$ is the matrix with all ones, m is the length of time series and $\mathbf{S}$ matrix is defined as $\mathbf{S} = \sum_{\mathbf{x}_i \in \mathbf{G}_e^{\mathbf{K}}} \left(\mathbf{x}_i \cdot \mathbf{x}_i^{\mathrm{T}}\right)$, $\mathbf{K} = \{I, II, III, IV\}$.

To start cluster analysis one has to provide the parameter for a predetermined number of clusters. For each time series set in $\mathbf{G} = \left\{\mathbf{G}_e^{I}, \mathbf{G}_e^{II}, \mathbf{G}_e^{III}, \mathbf{G}_e^{IV}\right\}$, k-Shape clustering has been carried out and optimal number of clusters has been verified using silhouette score $c_{\mathbf{G}} = (d_1 - d_2)/\max(d_1, d_2)$, where $d_1 = d_{\text{mean}}\left(\mathbf{x}_e^{(j)} - \{\hat{\mathbf{x}}_e \in cl_n\}\right)$ is the mean distance between a sample $\mathbf{x}_e^{(j)}$ and all other time series in the *next* nearest cluster $\{\hat{\mathbf{x}}_e \in cl_n\}$, $d_2 = d_{\text{mean}}\left(\mathbf{x}_e^{(j)} - \{\hat{\mathbf{x}}_e \in cl_i\}\right)$ is the mean distance between sample $\mathbf{x}_e^{(j)}$ and all other time series in the *same* cluster $\{\hat{\mathbf{x}}_e \in cl_i\}$. Figure 2 shows the results (z-normalized time series) of daily clusters for 30 minutes intervals during a day.



I.

II.

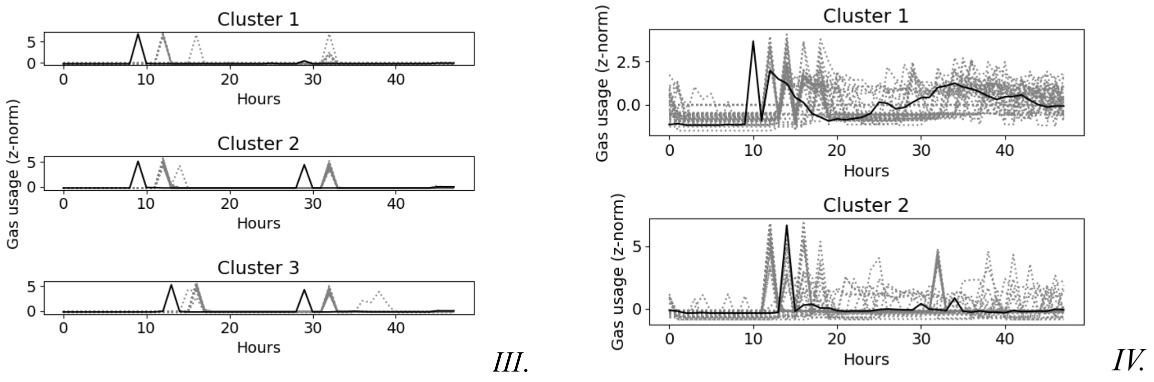


Fig. 2. Calculated daily clusters for each seasonal gas consumption (winter, spring, summer, autumn)

The calculated clusters for each set in $\mathbf{G}=\left\{\mathbf{G}_e^{I},\mathbf{G}_e^{II},\mathbf{G}_e^{III},\mathbf{G}_e^{IV}\right\}$ have discovered distinct peaks during the cold and transitional (autumn and spring) seasons, which can be generalized as $\left[\mathbf{G}_e\right]_{ma}$ (morning activities) and $\left[\mathbf{G}_e\right]_{aa}$ (afternoon activities). Also, the detected peak shifts in the vicinity of 8 a.m., 10 a.m., 4 p.m. and 8 p.m. (winter period) need more precise investigation, which will be covered in a separate research. The cluster analysis motivates us to using its results for gas disaggregation problem.

4. Approach to gas disaggregation problem. The identified peaks and activity windows $\left[\mathbf{G}_e\right]_{ma}$ and $\left[\mathbf{G}_e\right]_{aa}$ in section 3 can be used to extract non-heating components from aggregated gas time series. For this purpose, we propose the windowing algorithm to detect heating events by comparison of $x_e^{(j)}$ value at time j with $x_e^{(j+n)}$, i.e., if $\left\|x_e^{(j)}-x_e^{(j+n)}\right\|\le threshold$ occurs then heating event takes place. This should be done in the range of detected activity windows $\left[\mathbf{G}_e\right]_{ma}$ and $\left[\mathbf{G}_e\right]_{aa}$. To reinforce the proposed algorithm, the linear correlation between gas time series and outdoor temperature $\mathbf{I}_t$ can be taken into account. In addition, the pattern/signature extraction within activity windows is planned to be developed in further research.

References

1. Matallanas E. Neural network controller for Active Demand-Side Management with PV energy in the residential sector / E. Matallanas, M. Castillo-Cagigal, A. Gutiérrez. // Applied Energy. – 2012. – №91 (1). – P. 90–97.

2. Vamsikrishna P. Raspberry PI controlled SMS-Update-Notification (Sun) system / Vamsikrishna, Patchava; Sonti Dinesh Kumar; Shaik Riyaz Hussain; Rama Naidu, K. Proceeding of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT 2015), pp.1-4, 5-7 March 2015.

3. Kane T. Heating behaviour in English homes: An assessment of indirect calculation methods / T. Kane, S. Firth, T. Hassan, V. Dimitrou // Energy and Buildings. – 2017. – № 148. – P. 89–105.

4. Nornadiah R. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests / Razali Nornadiah, Yap Bee Wah // Journal of Statistical Modeling and Analytics. – 2011. – №2 (1). P. 21–33.