Viktória VÁMOS
László CZÉTÁNY
Miklós HORVÁTH
Tamás CSOKNYAI
Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Department of Building Services and Process Engineering

Gas Consumption Analysis for Educational Buildings

Analýza spotřeby plynu ve vzdělávacích budovách

A vast energy consumption database is available in the framework of the research project entitled "Large Scale Smart Meter Data Assessment for Energy Benchmarking and Occupant Behaviour Profile Development of Building Clusters". The database contains consumption data for approximately 10,000 buildings. Amongst the smart metered buildings, there are both residential and non-residential types. This research aims to identify different consumer groups and energy consumption profiles for various building types. For this study, a small sample was selected, which includes 76 school buildings. The energy consumption data are examined by using different clustering techniques: K-means, Fuzzy K-means, and Agglomerative Hierarchical Clustering Methods. In this article, the current state of our research is summarised.

Keywords: *energy consumption, cluster analysis, k-means clustering, fuzzy k-means clustering, hierarchical clustering*

V rámci výzkumného projektu s názvem „Posouzení monitorovaných dat ve velkém měřítku pro energetické srovnávání a vývoj profilu chování osob pobývajících ve skupině budov". Databáze spotřeby energie obsahuje údaje o spotřebě 10 000 budov. Mezi monitorovanými budovami jsou obytné i nebytové budovy. Cílem tohoto výzkumu je identifikovat různé skupiny spotřebitelů a profily spotřeby energie pro různé typy budov. Pro tuto studii byl vybrán malý vzorek, který zahrnuje 76 vzdělávacích budov. Údaje o spotřebě energie se zkoumal pomocí různých technik shlukování, jako: K-means, Fuzzy K-means a Aglomerative Hierarchical Clustering Methods. V tomto článku je shrnut současný stav probíhajícího výzkumu.

Klíčová slova: *spotřeba energie, shluková analýza, k-klastrování, fuzzy k-klastrování, hierarchické klastrování*

INTRODUCTION

Smart meter technology [1] is becoming more and more popular and is currently available in several buildings [2]. With the help of smart meters, vast amount of energy consumption data can be collected and analysed. A deeper knowledge of different energy consumption types – not only their amount, but their profile as well – can point out peak and off-peak periods in daily and longer-term energy use, and can provide essential information to better deal with demand side management (DSM). Some of the objectives of DSM are to balance the energy production and energy consumption, decrease the cost of energy, reduce the energy losses and improve the energy efficiency [3]. However, load shifting is the only DSM aspect which also includes a set of tools to improve the user's behaviour in saving energy. The development of appropriate DSM strategies requires an understanding of energy consumption and identification of the consumer group. In this research, a cluster analysis was used to examine the gas consumption of Hungarian school buildings, which is one of the objectives in the "Large Scale Smart Meter Data Assessment for Energy Benchmarking and Occupant Behaviour Profile Development of Building Clusters" research project. In the case of heating, it is, theoretically, an option to use heat storage to shift the load from the day to the night, but it would require having a large storage tank, thus, the storage losses would be significant as well, so this solution is not applied in practice. However, analysing the data consumption of the improper operation, behavioural problems can be detected and energy management can make actions to improve the situation. In addition to this, by examining the gas consumption of schools, buildings with a higher consumption can be filtered out.

METHODOLOGY

The methodology of our research is summarised in Fig. 1, which shows the flowchart of our study.

Data collection

The data collection was undertaken in the framework of the project mentioned above.

Data selection

The examination started with the data selection. The database contained empty periods and false, registered data; thus, these buildings were removed from the database. Overall, 76 school buildings with at least 1 year-long hourly gas consumption data were analysed.

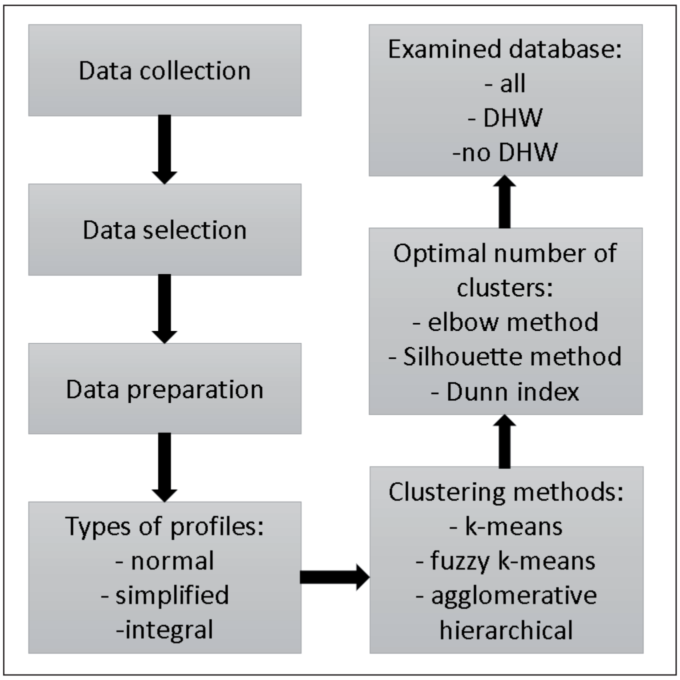


Figure 1 Flowchart of the examination process

Data preparation

In the data preparation step, the daily representative gas consumption profile was determined for each building based on the hourly state of the gas meters. The correction of the gas consumption is necessary if the heating demand is supplied with a gas boiler due to its dependence on the outdoor temperature. The available outdoor temperature database [4] - which contains detailed, hourly data for different locations - was defective for certain periods. The possibility of filling in the missing data was examined, but it could only be performed with an unacceptable level of accuracy. The original, full temperature database was compared with the refilled database - in which case, a database of the missing data, i.e., hour 1, 2, 3, … 23, was created and refilled by the mean value of the surrounding data – and the correlation coefficient between them was calculated. The acceptable value of this coefficient was determined to be at least 0.9. The weather in Hungary can greatly change in a short period of time, in such an extent that the level of accuracy in the correction is acceptable if only 1 hour of missing data is replaced. In our case, more data points were missing in the majority of places and, thus, the correction of the gas consumption based on the outdoor temperature could not been solved. The derogation caused by the weather conditions was taken into account by the non-dimensionalisation of the hourly data series by the total daily consumption for each day and each building. For each building, one typical daily consumption profile was created with an hourly resolution as the mean value of the given hourly consumption of all the examined weekdays.

Types of profiles

Three different profile types were compared to each other: a normal, hourly consumption profile, a simplified consumption profile and a daily integral consumption profile.

The simplified consumption profile was used by Yilmaz et. al [5] as well. Looking at a normal daily profile reveals that the maximum of the profile occurs in a single time interval, but rarely at the exact same time. So, by averaging the consumption data for periods longer than the sampling time (in this case longer than one hour), this effect can be reduced. In this case, periods of three hours long were used, and each point in the simplified profiles represents the mean hourly consumption for the three hours. At the x-axis, the middle of the interval is shown. In this way, the shape of the daily profiles become smoother, thus the nature of the consumer behaviour could be identified more clearly. In such a way, for example, the buildings, where there is a long heating up period before the start of education, would be classified into the same cluster, even though this period starts at 4 o'clock or 5 o'clock in the morning in some cases.

The integral profile is basically a cumulative representation of the normal profile. It is non-dimensionalised by the daily consumption. Fig 2 shows examples of the three different types of consumption profiles.

Clustering methods

To analyse the gas consumption data, three different clustering methods were used: k-means, fuzzy k-means and agglomerative hierarchical methods [6].

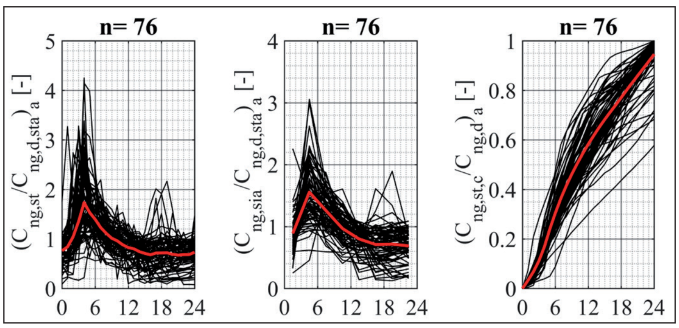


Figure 2 Examined data types (normal, simplified, integral)

The k-means clustering is a hard-clustering technique which clearly orders the consumption profiles into one cluster. At the beginning of the process, the number of clusters (*k*) has to be determined and k initial profiles need to be supplied as the centroids of the k clusters. After that, the distance between the m profiles and the k centroids has to be calculated. The profiles are ordered into the cluster from which the centroid's distance is a minimum. In the next step, the centroids of the clusters are calculated as the mean value of the profiles in the given cluster and the classification starts again. This iterative procedure lasts until the changes drop below a determined level.

In this research, the distance between the profiles and the centroids was determined by calculating the Euclidean distance. The final result of the k-means method depends on the first cluster centroid sets and, therefore, different cases should be examined and compared. In our paper, fifty different initial centroid sets were used: the result of the hierarchical clustering and another forty-nine sets were generated based on the available profiles. For each set, the silhouette was calculated, and the final result was chosen to maximise the silhouette. The iteration was repeated in every case until the difference between the centroids in two successive iterations became zero.

The fuzzy k-mean clustering method is a soft-clustering technique which means that the profiles are ordered not only to one cluster, but into every cluster with a determined probability. This probability is the membership degree calculated by the distance between the given profile and the cluster centroids. To determine the membership degree, the fuzziness parameter has to be given, which determines the fuzziness level of the clustering method. In our case, it was set to 1.5; however, in the literature, the usually recommended fuzziness value is 2 [7]. The reason is the following: when the recommended fuzziness parameter was used, the final cluster centroids for this dataset were too similar to each other and, therefore, the clustering became meaningless. The fuzzy k-means clustering calculation method is similar to the k-means clustering technique, but the cluster centroids are recalculated considering the membership degree. The iterative procedure was repeated, in this case, until the changes in the cluster centroids dropped below 1% in two successive iterations.

The agglomerative hierarchical clustering method works in steps and there is no need for any iteration in this case. In the first step, every profile is in a different cluster, so the number of the cluster is equal to the number of the profiles and their centroids are the consumption profiles. There are several types of solutions to represent the clusters and calculate the distance between them to be able to determine which ones should be merged. In this paper, the distance between each cluster's centroids are calculated and the two nearest clusters are merged. The centroid of the merged cluster is the mean value of the profiles belonging to it. For this clustering method, there is no need to predetermine the number of clusters, only the calculation has to be stopped at the required level.

Each parameter influences the final results: the used distance formula, the initial cluster centroids and the number of iterations in the case of k-means and fuzzy k-means methods and the fuzziness parameter in the fuzzy k-means method. The effect of these parameters has been analysed before [7], [8] but more investigation is required.

Optimal number of clusters

To determine the optimal number of clusters, different measures can be used. In this paper, the elbow [9] and silhouette methods were used and Dunn's index was calculated [10]. Using the elbow method, the distances between the profiles and the centroids of their clusters have to be summarised and plotted according to the number of clusters. The 'elbow' of this line presents the optimal number of clusters. The silhouette method determines the silhouette value, which measures the goodness of the clustering result. The closer the profiles are in the same cluster and the further the profiles are in the different clusters, the better the results are. The optimal number of clusters is at the maximum silhouette value. The idea of the Dunn index is similar to the silhouette value calculation, but it measures the maximum distance in a cluster and the minimum distance between the different clusters. The maximum Dunn index indicates the optimal number of clusters. In the literature, there are several different calculation methods of Dunn's indexes, but only one of them was used in this research.

Examined database

The school buildings were divided into two groups depending on their hot water production type and their consumption was analysed for two different time periods. The time periods are the 'summer' period from May to September when there is, most likely, no heating and the 'winter' period from October to April when heating is necessary. First the 'winter' gas consumption of all the buildings ('all') was examined. Then the buildings were divided into two parts according to their 'summer' consumption. In the process of deviation, nine different cases were examined: 3 types of data X 3 types of clustering methods. The buildings that hardly had any gas consumption in the summer were in the 'no DHW' buildings at least 5 out of 9 times. In these buildings, no gas equipment was used other than for reasons of heating. The other group is 'DHW' buildings, where gas boilers were used for heating domestic hot water (DHW) or gas ovens were used for cooking. The applied clustering methods are particularly useful if we do not know the building, but just have the monitoring data. Therefore, the division into 'DHW' and 'no DHW' clusters was based purely on the consumption data analysis. In the case of the 'DHW' buildings, the 'summer' and the 'winter' consumption were both calculated, but in the case of the 'no DHW' buildings, only the 'winter' period was examined.

COMPARISON OF DIFFERENT OPTIONS

Optimal profile type

First, the impact of the profile types was examined. Tab. 1 shows the fuzzy k-means clustering results of 'all' the buildings for the 'winter' season. The results were calculated for three clusters with different profile types. The results for the normal profiles were selected as the reference and the simplified and integral data were compared to that. The buildings were colour coded in the normal profile type clustering, thus, each cluster had a specific colour and the buildings' colour code was not changed for the simplified and integral profile type clustering. Ideally, for all the profile types, the clustering results would be the same, but Tab. 1. proves that the type of the profile affects the result of the clustering. It is visible that both the classification of the consumption profiles and the size of the clusters vary as well. It is essential to choose the profile type carefully: despite the cluster centroids – the shape of the representative profiles – could be similar to each other (Fig. 3 and Fig. 4), where different buildings are clustered into the same group. From the DSM point of view, it means that the methodological choice affects the consumer group in which the given building shall be classified into and, as a consequence, the applied tariff as well.

Optimal clustering method

The clustering method plays the most critical role in the clustering process and has the most significant effect on the results. The hierarchical clustering technique proved to be improper for these buildings. Because of its nature, irregular consumption profiles were first selected and classified into separate clusters. However, the large groups of buildings, which otherwise should be divided (by both visual inspection and according to the k-means or fuzzy k-means clustering) into more groups, stayed in the same group. So, this kind of clustering method is very sensitive to incorrect data and could not separate the other buildings effectively.

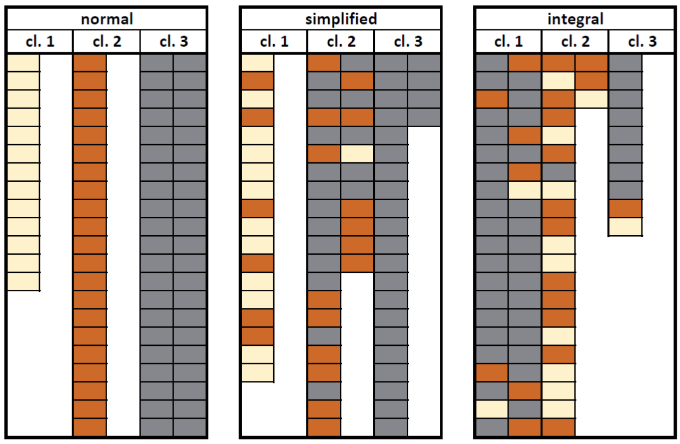
The result of the k–means clustering is not as consistent and repeatable as the hierarchical method because it iterates several times and its final result depends on the initial centroids. Still, it is also sensitive to the irregular data, if the number of clusters is higher. This method was modified to eliminate the problem of one profile cluster: only clusters with at least 3 profiles in them were accepted. However, the results were not acceptable because the shape of the final cluster centroids were very different compared to the results given by the other clustering methods, thus, this idea was rejected.

Optimal number of clusters

Fig. 5. shows the value of the different indexes by the different methods for the selection of the optimal number of clusters. In the case of the silhouette method and the Dunn index calculation, the maximum values objectively show the optimal number of clusters. However, in the case of the elbow method, the 'elbow' of the line has to be found by visual inspection. It is visible that the value of the optimal number of clusters is influenced by the used clustering method and index. Therefore, using the different techniques, different consumer groups could be identified.

The optimal number of the clusters applying the various different methods for the different databases ('all' buildings 'winter' period, 'DHW' buildings 'summer' and 'winter' period, 'no DHW' buildings 'winter' period) was determined and the final clusters were examined. In every case, the optimal result was chosen between the three cluster index options. Even though the result could not be determined objectively using the elbow method, this method appeared to be the most appropriate.

Table 1 Fuzzy k-means clustering results for 3 clusters



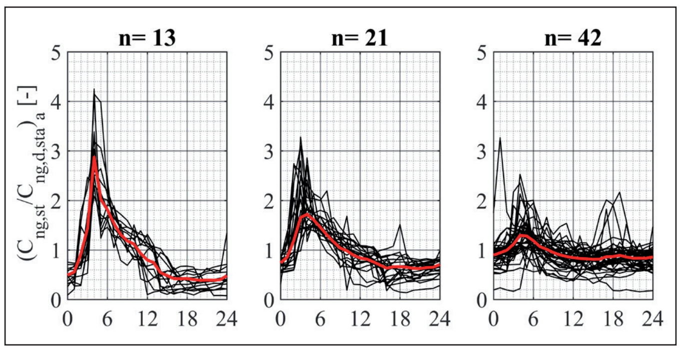


Figure.3 Clustering results with normal data

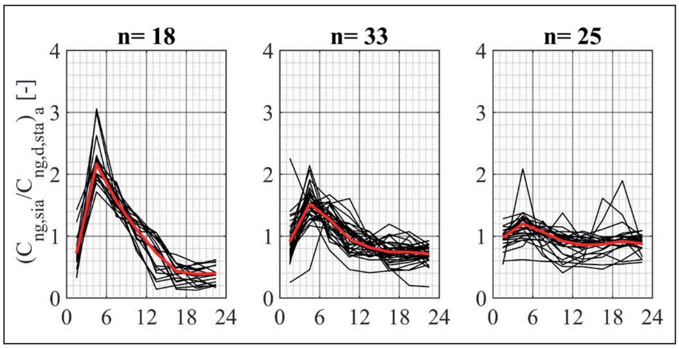


Figure 4 Clustering result with simplified data

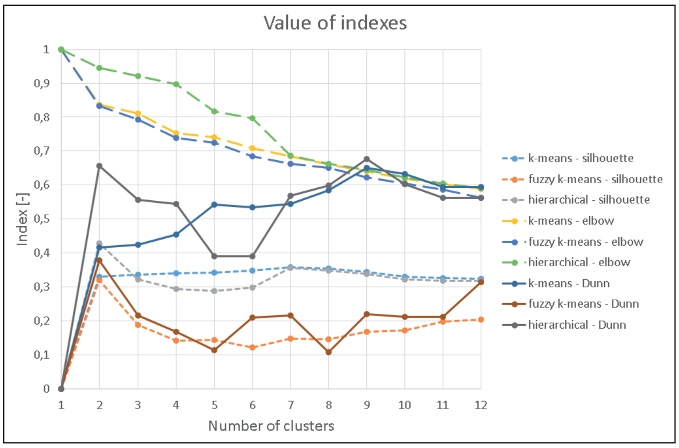


Figure 5 The value of the different indexes in all the buildings in the 'winter' season

So, in conclusion, for the normal profiles, the fuzzy k-mean clustering method and the elbow method were the best fit for the purpose. These results are selected to identify the characteristic energy consumption profiles.

RESULTS

The 'winter' period was examined for three different groups of buildings: 'all' the buildings, the 'DHW' buildings and 'no DHW' buildings. The optimal number of clusters was 2 in the case of 'all' the buildings (Fig. 6), 4 in the case of the 'DHW' buildings (Fig. 7) and 3 in the case of 'no DHW' buildings (Fig. 8). In every case, there is a heating up period before the education starts, only the amplitude and the length of this period changes.

In case of the 'DHW' buildings, one cluster shows another high peak in the evening. In this case, it is reasonable to assume that the buildings belonging to this cluster are schools with dormitories. In these buildings, the students consume domestic hot water for showers and/or to warm up their rooms before sleeping. The figures also show that the gas consumption is not reduced below the afternoon level in many buildings during the unoccupied hours. This behaviour can lead to high energy losses.

The 'summer' (Fig. 9) and 'winter' (Fig. 7) period consumption of the 'DHW' buildings were compared. The 'summer' consumption seems to be higher than the 'winter' consumption, but these are non-dimensional values. Small consumptions during a low consumption period could result in a high relative value. During the 'summer' period, the heat up term can be observed in one cluster only. These buildings either have huge heat losses and some heating is required in May or September or

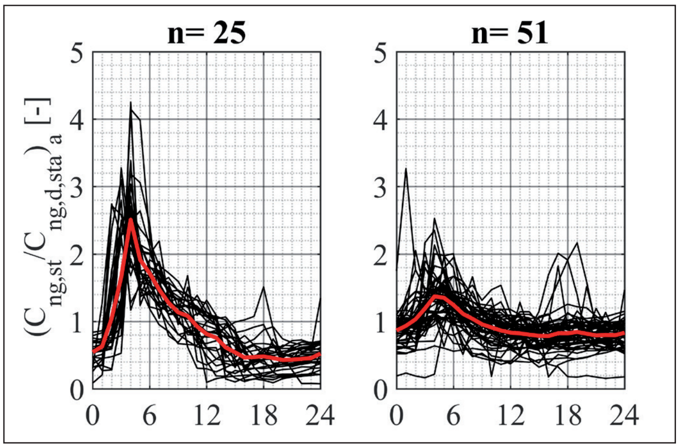


Figure 6 Clustering results of 'all' the buildings in the 'winter' period

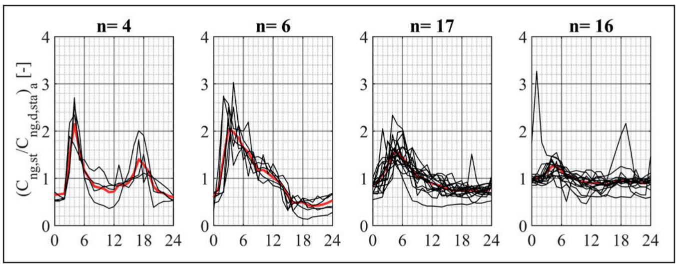


Figure 7 Clustering results of the 'DHW buildings in the 'winter' period

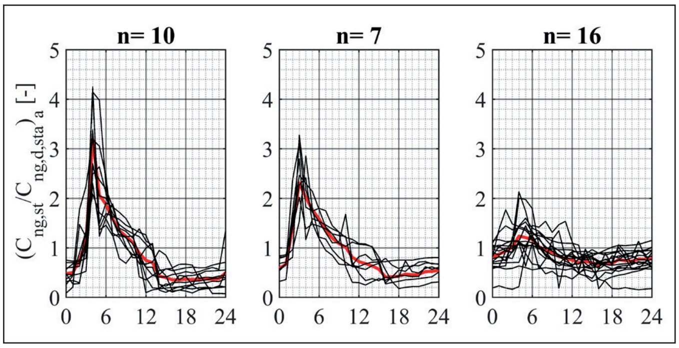


Figure 8 Clustering results of the 'no DHW' buildings in the 'winter' period

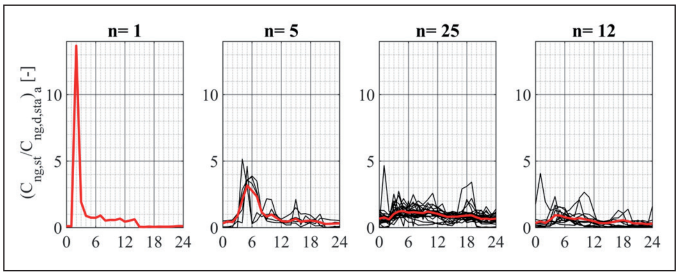


Figure 9 Clustering results of the 'DHW' buildings in the 'summer' period

the DHW is prepared during this off-peak period and is stored for further daily use. One cluster contains one consumption profile only, which can be an incorrect one and that is the reason why it is selected separately.

CONCLUSION

Despite the limited number of educational buildings with appropriate consumption data, many different options were examined and compared.

First, the type of the data was analysed. The type of the data is very important and effects the results of the clustering methods. Small modifications can lead to significant changes. The integral profile type is inadequate to express the occupant behaviour and control strategy. Therefore, it is not recommended to be used in the process of a daily gas consumption profile determination.

Second, different clustering methods were compared. From the examined methods, the fuzzy k-means technique was the most appropriate because it could handle the incorrect data as well. Despite the recommendation, β =1.5 was selected as the fuzziness parameter. The reason given is that if the recommended value of β =2 was used, the final cluster centroids were too similar to each other. The mathematical background is: if the value of the fuzziness parameter is higher, all the profiles have a larger influence on all the cluster centroids.

Third, the optimal number of clusters was searched. To determine the optimal number of clusters, the elbow method was the most useful despite the fact that the results of this method could not be determined objectively.

In conclusion, the most accurate clustering results are obtained if the normal data type was used and the fuzzy k-means clustering method was applied. The optimal number of clusters could be determined most reliably with the elbow method.

Then, the obtained typical profiles were observed to find explanations in the building's operation on the shapes of the graphs. During the 'winter' period, the heating up interval before the start of lessons can always be observed. In one cluster, a high afternoon/night/evening peak appears as well. The reason is that the buildings that belong to this cluster are schools with dormitories. The lack of a setback operation could also be observed in many cases. It means that the heating system still serves the building during the unoccupied periods. This inappropriate operation leads to higher costs and the loss of energy. The 'summer' and 'winter' period results were also compared. The heat up period in the buildings (except for one cluster) disappears in the 'summer'. Only some of the buildings require heating during the transition period and these schools may have high heat losses.

The proposed methodology opens new perspectives in a building's operation. Analysing the energy consumption results, conclusions could me made about the buildings' operation and the consumers' behaviour. The energy-wasting buildings could be filtered out, the gas consumer equipment and the physical status of the building could be predicted without knowing the examined building. The knowledge of any kind of energy consumption could help decision makers to develop effective DSM strategies and determine energy tariffs which encourage people to save energy.

The main results of this research will continue to be used in our future work. We intend to examine different kinds of energy consumption and different types of buildings. The representative consumption profiles of each building type would help to develop effective operation strategies. The energy-saving potential of the Hungarian buildings could be calculated and compared with the national data.

Acknowledgment: The work has been carried out within the research project entitled "Large Scale Smart Meter Data Assessment for Energy Benchmarking and Occupant Behaviour Profile Development of Building Clusters". The project (no. K 128199) has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K_18 funding scheme.

The research reported in this paper and carried out at BME has been supported by the NRDI Fund (TKP2020 IES, Grant No. BME-IE-MISC) based on the charter of bolster issued by the NRDI Office under the auspices of the Ministry for Innovation and Technology.

Contact: vamos@epget.bme.hu

References

[1] D. B. AVANCINI, J. J. P. C. RODRIGUES, S. G. B. MARTINS, R. A. L. RABÊLO, J. AL-MUHTADI, P. SOLIC. Energy meters evolution in smart grids: A review. *J. Clean. Prod.* 2019, vol. 217, pp. 702–715. doi: 10.1016/j.jclepro.2019.01.229.

[2] J. LEIVA, A. PALACIOS, J. A. AGUADO. Smart metering trends, implications and necessities: A policy review. *Renew. Sustain. Energy Rev.* 2016, vol. 55, pp. 227–233. doi: 10.1016/j.rser.2015.11.002.

[3] A. F. MEYABADI and M. H. DEIHIMI. A review of demand-side management: Reconsidering theoretical framework. *Renew. Sustain. Energy Rev.* 2017, vol. 80, no. March, pp. 367–379. doi: 10.1016/j.rser.2017.05.207.

[4] Ogimet Weather Information Service [online]. 2018. http://ogimet.com/index.phtml.en.

[5] S. YILMAZ, J. CHAMBERS, M. K. PATEL. Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management. *Energy*. 2019, vol. 180, pp. 665–677. doi: 10.1016/j.energy.2019.05.124.

[6] G. CHICCO, R. NAPOLI, F. PIGLIONE. Comparisons Among Clustering Techniques for Electricity Customer Classificatio. *IEEE Trans. POWER Syst.* 2006, vol. 21, no. 2, pp. 933–940.

[7] K. Le ZHOU, C. FU, S. L. YANG. Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation. *Sci. China Inf. Sci.* 2014, vol. 57, no. 11, pp. 1–8. doi: 10.1007/s11432-014-5146-0.

[8] A. SINGH, A. YADAV, A. RANA. K-means with Three different Distance Metrics. *Int. J. Comput. Appl.* 2013, vol. 67, no. 10, pp. 13–17. doi: 10.5120/11430-6785.

[9] I. DENT, U. AICKELIN, T. RODDEN. Application of a Clustering Framework to UK Domestic Electricity Data. *SSRN Electron. J.* 2017. doi: 10.2139//ssrn.2829232.

[10] O. ARBELAITZ, I. GURRUTXAGA, J. MUGUERZA, J. M. PÉREZ, I. PERONA. An extensive comparative study of cluster validity indices. *Pattern Recognit.* 2013, vol. 46, no. 1, pp. 243–256. doi: 10.1016/j.patcog.2012.07.021.

Nomenclature

DHW	domestic hot water
k	number of clusters
m	number of profiles
n	number of profiles in the same cluster
β	fuzziness parameter
$C_{ng,st}$	the amount of natural gas consumed between two sampling intervals [m^3/h]
$C_{ng,d,sta}$	the amount of gas consumed during the specific day divided by the number of samplings per day [m^3/h]
$C_{ng,st,c}$	the cumulative distribution function of the natural gas consumed for the specific day [m^3]
$C_{ng,d}$	the amount of gas consumed during the specific day [m^3]
$C_{ng,sia}$	the amount of natural gas consumed between two intervals used to construct the simplified profile divided by the samplings per simplified intervals [m^3/h]
subscript "a"	the average of the profiles investigated over the whole year