Energy & Buildings 252 (2021) 111376

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

Development of electricity consumption profiles of residential buildings based on smart meter data clustering

László Czétány [a], Viktória Vámos [a], Miklós Horváth [a,*], Zsuzsa Szalay [b], Adrián Mota-Babiloni [c], Zsófia Deme-Bélafi [a], Tamás Csoknyai [a]

[a] Budapest University of Technology and Economics, Faculty of Mechanical Engineering, Department of Building Services and Process Engineering, Műegyetem rkp 3, Budapest 1111, Hungary
[b] Budapest University of Technology and Economics, Faculty of Civil Engineering, Department of Construction Materials and Technologies, Műegyetem rkp 3, Budapest 1111, Hungary
[c] ISTENER Research Group, Department of Mechanical Engineering and Construction, Universitat Jaume I (UJI), Castelló de la Plana E-12071, Spain

ARTICLE INFO

Article history:
Received 18 May 2021
Revised 23 July 2021
Accepted 18 August 2021
Available online 21 August 2021

Keywords:
Electricity consumption profile
Smart meter
Data clustering
K-means
Fuzzy k-means
Hierarchical
Residential buildings

ABSTRACT

In the present research, a high-resolution, detailed electric load dataset was assessed, collected by smart meters from nearly a thousand households in Hungary, many of them single-family houses. The objective was to evaluate this database in detail to determine energy consumption profiles from time series of daily and annual electric load. After representativity check of dataset daily and annual energy consumption profiles were developed, applying three different clustering methods (k-means, fuzzy k-means, agglomerative hierarchical) and three different cluster validity indexes (elbow method, silhouette method, Dunn index) in MATLAB environment. The best clustering method for our examination proved to be the k-means clustering technique. Analyses were carried out to identify different consumer groups, as well as to clarify the impact of specific parameters such as meter type in the housing unit (e.g. peak, off-peak meter), day of the week (e.g. weekend, weekday), seasonality, geographical location, settlement type and housing type (single-family house, flat, age class of the building). Furthermore, four electric user profile types were proposed, which can be used for building energy demand simulation, summer heat load and winter heating demand calculation.

1. Introduction

1.1. Objective

Building energy regulations focus on increasing the performance of building structures and energy supply systems. This way, overall energy use associated with building characteristics is decreasing. However, it was observed that human behaviour in buildings plays an essential role in the energy balance of a building [1]. When an occupant turns on the heating, opens the window or switches on the light, the building's energy balance affects the overall energy consumption. The rebound effect makes the issue even more critical in energy-efficient buildings: as operation costs become significantly lower, users lost an important motivation factor in saving energy. Significant effort has been made recently (largest international cooperative project: IEA EBC Annex 66 [2]) to investigate and understand building occupants' everyday activities, presence schedules and overall attitude to sustainability that influence energy consumption.

A principal research trend in occupant behaviour modelling is the application of data-driven methods [3,4]. In the past, only some segments of the building stock's and building users' energy performance could be analysed simply because of the lack of detailed consumption-related information. Nowadays, the most significant challenge in this area of research is still the lack of high-resolution data on a large building population.

In our research, a high-resolution, detailed electric load dataset was used which was collected by smart meters in nearly a thousand households in Hungary, many of them single-family houses. Our objective is to evaluate this database in detail, determine energy consumption profiles from time series of daily and annual electric load, and examine the factors influencing the profiles. There are only a few examples in the literature to assess a building sample of this size. We do not know of any similar analysis in the Eastern European region, which is important because the economic and cultural environment affects consumer habits as well.

* Corresponding author.
E-mail address: horvath@epgep.bme.hu (M. Horváth).

https://doi.org/10.1016/j.enbuild.2021.111376
0378-7788/

Nomenclature

Abbreviations

m_1-12	data of months 1 (January) to 12 (December) used for creating profiles
m_4	data of month 4 (April) used for creating profiles
d_1-7	data of days 1 (Sunday) to 7 (Saturday) used for creating profiles
d_3;4	data of days 3 (Tuesday) and 4 (Wednesday) used for creating profiles
m_1_d_1;7	data of days 1 (Sunday) and 7 (Saturday) from month 1 (January) is used for creating profiles
DHW	domestic hot water

Latin letters with different subscripts

$d(p_n,p_m)$	Euclidean distance between the n^{th} and m^{th} profile examined [-]
N	number of data points in the profile type chosen (N = 97 for daily profiles with sampling time 15 min and N = 12 for monthly profiles) [-]
MD(n,k)	membership degree of profile n related to cluster k [-]
β	fuzziness parameter (here 1.5 is used) [-]
a_i	cohesion of profile examined (def. as the average distance between profile and other profiles in the same cluster) [-]
b_i	separation of profile examined (def. as the average distance between profile and other profiles in the kth cluster, K-1 pieces can be calculated from it, and the minimum values are chosen) [-]
S_i	silhouette score (or width) calculated for the profile [-]
$P_{a,st}$	average of electric power consumption for sampling time [kW]
$P_{ya,st}$	whole year average of electric power consumption for the sampling time [kW]
$C_{e,da}$	daily electricity consumption averaged for the month [kWh]
$C_{e,y}$	daily average electricity consumption for the year [kWh]
subscript a	average -

1.2. Consumption profiles

Electricity consumption profiles follow daily, weekly and seasonal trends. Their shape is strongly influenced by the presence and activity of the consumer and the characteristics of the equipment used.

Detailed energy consumption profiles can be utilised by energy supply and utility companies to forecast their energy (annual, monthly, weekly and daily) production and supply. Comparative analysis of electric production daily trends and demand-side profiles can provide essential information for electric grid operators giving peak-shifting opportunities, decreasing operating hours of peak power plants.

Characteristic consumer profiles can be recognised purely from the database using clustering procedures. Such methods were also used in the present research [5–7].

With the help of more accurate and up-to-date consumer profiles, more accurate calculations can be made with dynamic building simulation. Therefore, we've also aimed to develop profiles that can be used for simulation practice.

1.3. Research question

Based on the preliminary data filtering, we narrowed our research to 649 residential units (60.2% of them are family houses). For this population, we determined daily electricity consumption profiles and examined how the following factors influence the shape of the profiles:

- type of meter and electric circuit in the housing unit (e.g. peak, off-peak meter);
- day of the week (e.g. weekend, weekday);
- effect of seasonality on daily profiles;
- geographical location, type of settlement;
- type of building unit (single-family house, flat, age class of the building).

The limited information available on the building units and the measured data narrowed the research possibilities to the listed factors.

We also examined the evolution of the annual profiles to see which months can be characterised by higher consumption during the year.

Finally, we proposed electrical profiles for dynamic simulation modelling considering the specific data requirements of the simulation. Although representativity was not a criteria when the meters were installed (authors had no impact on the site selection), we proved that the analysis can provide statistically acceptable results for Hungary (see table 4 and related text).

The paper is organised as follows: in Section 2 the theoretical and calculation background is presented. This includes analysed smart meter data and the building surveys made during the evaluation. During the research the k-means, fuzzy k-means and hierarchical clustering techniques were tested and evaluated by using the elbow and the silhouette methods and calculating the Dunn index. In Section 3 the main results are presented based on the analysis, which include daily and yearly electricity load profiles and their evaluation. In Section 4 the main conclusions are presented along with proposed electric load profiles for residential buildings, which can be used for modelling household equipment heat load for heating and cooling system sizing and yearly building energy simulation.

2. Theory/calculation

2.1. Smart meter data

2.1.1. Background

In December 2018, the European Union (EU) revised the Renewable Energy Directive. They targeted to 32% the overall Renewable Energy Sources consumption by 2030 [8]. There, contrary to the first approved Renewable Energy Directive [9], the relevance of the Smart Cities and Smart Communities was highlighted to stimulate the development of renewable energy and energy efficiency. Similarly, the same year, the Amendment [10] to the Energy Performance of Buildings Directive [11] considered integrating renewables to smart grids and the smart readiness indicator part of the digitalisation of the building sector. According to this Directive, smart-ready systems can save energy by providing more accurate information about consumers' consumption patterns and enabling a more effectively grid management and highlighting the relevance

of interoperability between their systems, including smart meters. These points were also highlighted by Aubel and Poll [12], who added that the cost and hassle of the meter readings are reduced, and the possibility of fraud is reduced.

Mainly, smart meters could increase renewable power in new niche energy markets [13]. Therefore, smart meters represent a key factor for energy saving in the EU buildings, responsible for approximately 40% of energy consumption and 36% of CO_2 emissions. According to the European Commission's report from 2014, close to 200 million smart meters for electricity and 45 million for gas would have been present by 2020 in the Member States. By 2020, it was expected that almost 72% of European consumers should have a smart meter for electricity, while 40% will have one for gas [14]. However, the EU aimed to replace at least 80% of electricity meters with smart meters in households and commercial buildings by 2020, wherever it is cost-effective to do so [15]. Currently, one can find significant differences in the progress of the smart meter deployment in the Member States, the roll-out of smart meters in Spain, Italy, Sweden, Estonia, and Finland is above 80% whereas, for Germany, Ireland, Croatia, Greece, and the Czech Republic, no smart meters are introduced [16].

All in all, the EU is considered a reference in the smart meter roll-out by other non-EU countries, like Brazil [17]. Zhou and Brown performed a comparative study of five European countries. They proved that government interventions highly influence smart meter deployment and that these systems do not necessarily diffuse more rapidly in countries with a more robust manufacturing capability [18]. Similarly, Hielscher and Kivimaa confirmed that the smart meter evolution is strongly connected to UK energy policies [19].

Information gathered by smart meters and other advanced metering infrastructures can be used in several ways. According to the review of Kabalci, most of the current studies about smart metering aim to increase energy efficiency, demand management, utility planning, cost control, and constructions [20]. For instance, Razavi et al. used electricity consumption data from more than 5 000 households to predict the present and future home-occupancy status. They correlated the results with various household and building characteristics [21]. Kiguchi et al. based their research on 646 Irish households for predicting the intra-day load profiles. They developed a statistical model that can forecast the impact of a Time-of-use tariff considering lifestyle constraints [22]. Furthermore, smart sensing systems could also enable new opportunities for making real-time decisions [23].

Notable efforts have been made in the use of energy meter daily profiles in commercial building simulation. Roach proposed a multimodel inference approach to prove the influence of tenant and equipment behaviour in 129 commercial buildings' seasonal electricity demand [24]. Li et al. carried out time-domain and frequency-domain analyses on smart meter data of 188 commercial office buildings to extract and quantify distributions of key load profile parameters and periodic fluctuations and load variability, respectively [25]. Najafi et al. reduced from 290 to 29 the features utilized for estimating the buildings' use type and improved accuracy from 71% to 74%. Then, when reduced from 224 to 17 features, the performance class classification accuracy increased from 56% to 62% [26].

Zhu et al. applied an effective framework based on three steps: simple and efficient algorithms preprocess the high-frequency building load time series into a set of meaningful daily profiles, then, selecting appropriate data mining algorithms for appropriate prediction models, and finally, their residuals are analyzed by statistical quality control theory. For each load profile, a control chart with upper control limit is created.[27]. Gunay et al. developed an electricity end-use disaggregation method and tested it using data from an academic office building with six different load disaggregation scenarios. This method utilizes building automation system data to disaggregate low-frequency electricity data into major commercial building end-uses. Low-frequency meter data was disaggregated at a reasonable accuracy with the contextual information provided by the data [28]. Samadi and Fattahi quantify Energy Use Intensity for different load categories through a energy disaggregation model for institutional buildings. This model, that analises the relation between load categories and dependency factors, is able to distinguish daily-used devices and common loads deploying their dependency on workdays and occupancy factors [29].

Apart from the previous examples, in recent years, smart meter measurements are being used for detailed analysis and forecasting of household loads, clustering methods and classifying of different types of load profiles [30]. In this sense, Yildiz et al. provided guidelines to enable more objective comparisons between different models and studies [30]. Moreover, Torriti stated that electricity loads depend predominantly on the timing of human activities rather than prices and are easily predictable in offices than in residential buildings. Therefore, occupants and appliances characteristics will be a key factor, together with dwelling characteristics [31].

Ndiaye and Gabriel used data gathered from energy audits, phone surveys, and smart meter readings in a principal component analysis to generate household electricity consumption regression models [32]. They have reduced the number of variables from 59 to 9 to avoid multi-collinearities: the number of occupants, ownership, number of weeks of vacation per year, type of fuel used in the pool and domestic hot water heater, and heating system, the existence of air conditioning system and its typology, and number of air changes per hour (at 50 Pa). McLoughlin et al. obtained that household composition, number of bedrooms, water heating, and cooking type were the most influential variables on maximum household electricity consumption [33]. Kavousian et al. classified factors that influence household electricity consumption in major groups: external conditions, physical characteristics of dwellings, appliance, electronics stock, and occupants; being the first and second groups the most influential on the resulting hourly value [34]. Beckel et al. have designed and developed a system that classifies private households according to pre-specified properties [35]. The evaluation of 3 488 households proved that this system could classify with good accuracy eight properties: marital status, age of the building, members in the family, number of children, type of cooking facility, retirement, and employment of chief income earner. Tong et al. used smart meter readings from more than 5 000 Irish households to identify energy behaviour indicators through a cross-domain feature selection and coding approach [36]. Their results show that employment status and internet usage are highly correlated with household energy behaviour. Huebner et al. used data from 845 households have found that appliance ownership, and usage and household size are the most influential variables in understanding electricity consumption in natural gas-fuelled, centrally heated buildings [37].

Viegas et al. combined smart metering and survey data to classify new residential electricity customers using model-based feature selection [38]. They have proved that survey data increases up to 20% the accuracy of the classification. Using one week of metering data, more than 50% of the customers are correctly classified into four consumption groups. In the same way, Gouveia and Seixas used socio-economic characterisation of the household members (mainly number of occupants and monthly income), dwellings physical characteristics (year of construction and total floor area) and electrical heating/cooling equipment and fireplaces information (municipality of Évora, Portugal) to reduce the initial proposal of 10 clusters to 4, following 77% of them the U shape electricity annual profile [39]. Then, Gouveia and Seixas state that

hourly electricity consumption deviations from average for different maximum and minimum daily temperatures can be used to cluster households with comparable space heating and cooling behaviour [40]. Laicane et al. have classified households into 7 groups according to heating used in households [41]. After eliminating predictors with low significance, they saw that it is possible to explain part of the variability of electricity use and that the most statistically significant factors affecting household electricity use are the number of appliances and household floor area, followed by net income, year of construction and temperature set during winter. Dane and Swan have included domestic hot water and compared two smart meter electricity datasets [42]. The seasonal and daily observations prove that appliances, lighting, and plug-loads mainly drive other consumption profiles, but they recommend shorter time steps (5 min) to identify shorter duration loads.

McLoughlin et al. have continued their previous work and evaluated k-means, k-medoid, and SOM clustering methods, being the last one the most suitable [43]. Hache et al. have identified the transparent and straightforward set of socio-economic, dwelling and regional characteristics as critical drivers of the different levels of French households' energy consumption using the CHAID (Chi-Square Automatic Interaction Detection) clustering methodology [44]. Income is crucial, but the age of the household's representative, the family type and the tenure are factors that should be accounted for in determining the households' energy consumption. Azaza and Wallin used the electricity consumption variance to identify consumer groups (behaviour variability or customers' behaviour changes) with higher contribution to utility system peak, based on hierarchical and SOM clustering techniques [45]. They distinguished five clusters with marked characteristics. Khan et al. have proposed a clustering algorithm for big-data analysis of highly non-linear smart meter data profiles using a set of weighted linear profiles, considering different clustering scenarios [46]. The method extracts patterns with high intra-cluster pattern similarity, among other benefits. Funde et al. have developed and tested (on two smart meter datasets) a motif-based association rule mining procedure to determine the energy usage daily profile and to identify the association between the energy-consuming appliances [47].

Using a different kind of clustering methods to analyse the energy consumption data is a generally applied solution. No best clustering technique exists; their practicability depends on the input parameters. Therefore, several types of them have applied in energy consumption related studies. For example, in the work of Wang et al. GMM clustering method is used to identify the groups of district heating users based on their consumption pattern features [48]. Yilmaz et al. have applied the commonly used k-means clustering technique to analyse the electricity consumption data of Swiss households [49]. The electricity consumption profile of Chinese residential users was examined by the fuzzy c-means clustering method by Zhou et al. [50]. Li et al. have a new strategy proposed, which combines agglomerative hierarchical clustering technique with other methods to forecast the electricity consumption [51].

Using clustering techniques, several types of energy data could and were examined effectively. For example, in the work of Wang et al., the spatial hierarchical clustering method was applied to explore the geographical characteristics of the final energy consumption in China [52]. Pieri et al. have used the k-means clustering method to cluster Attica hotels based on energy consumption data and physical parameters [53]. This way the energy efficiency of the hotels could be compared with others with similar characteristics. The focus of Gianniou et al. was the daily heating consumption analysation of Danish single-family houses [54]. The k-means clustering method was used to determine the heating consumption profiles of district heating users, and the effect of building and occupant related characteristics was also examined. In addition, clustering techniques were applied to analyse the natural gas consumption in Algeria [55], water consumption in Greece [56] and occupancy data in UK [57].

The exact knowledge of electricity load patterns is important for many reasons. Various Demand Side Management (DSM) techniques can ensure the required energy needs economically and environmentally-friendly [58]. Applying these techniques, the knowledge of the energy consumption profile is required to develop the proper DSM method and ensure a balance between energy production and energy consumption. The energy consumption profiles are also essential input parameters for dynamic energy simulation models, and they are necessary for technical building system planning.

In this research, the electricity consumption data of residential buildings were analysed. In the literature, several examples could be found for similar investigations. For example, Stre-Meloy has examined the residential electricity usage of weekdays during the evening peak period [59]. As a result, the consumption profiles were sorted into two clusters, and the key influencing activity factors of classification were determined. Another example is from China [60]. In this research, an improved fuzzy c-means clustering method was used to determine characteristic monthly electricity consumption profiles based on December 2014. Wen et al. have developed an improved k-means clustering technique, and a shape-based method was suggested for examining residential electricity consumption data [61]. The improved k-means clustering technique was tested on two different datasets (from Ireland and China), the shape-based method was tested in the Irish dataset. Bourdeau et al. have investigated the electricity consumption data of higher education buildings [7]. For this analysis, three different k-means clustering techniques were examined and compared while the timeframes and time-steps of input data were modified.

2.1.2. Smart meter dataset and building survey

In a national demonstration project conducted by KOM Ltd., 128 634 smart meters were installed in different regions of Hungary (Fig. 1) [62]. The main goal of the demonstration project was to install meters in as many buildings as possible of different types (residential, public, commercial, and industrial) and different settlements (capital, cities, towns, and villages). However, statistical representativeness was not targeted. Meters were primarily installed in Central Hungary and the Southern Great Plains. While there were many smart meters installed in Northern Hungary and the Northern Great Plains, they were only in two cities (Nyíregyháza and Miskolc). Altogether 28 993 m measured electricity consumption, of which 24 917 were installed in residential buildings (Table 1).

KOM Ltd. provided us the anonymised data for research purposes within the framework of a bilateral cooperation agreement. Authors are not allowed to make the data publicly available Currently the dataset is managed by MVM ESCO Ltd. company, legal successor of KOM Ltd.

One of our research goals was to analyse the possible differences between building types and settlement types, so further elaboration of the datasets was necessary (Table 2). The electricity supplier provided us with the exact address only for 9 237 m. In the next step, preliminary filtering was applied to remove unreliable datasets (see 2.1.3), which resulted in 4 454 useful datasets. The buildings corresponding to these addresses were then surveyed with the help of a GIS mapping tool to gather qualitative information that may be relevant for further analysis. An expert identified the building and assessed the building function, building type, covered area, number of stories, general condition of the building, visible retrofit measures (change of windows, additional insulation on façade), type of roof (flat roof, pitched roof occupied

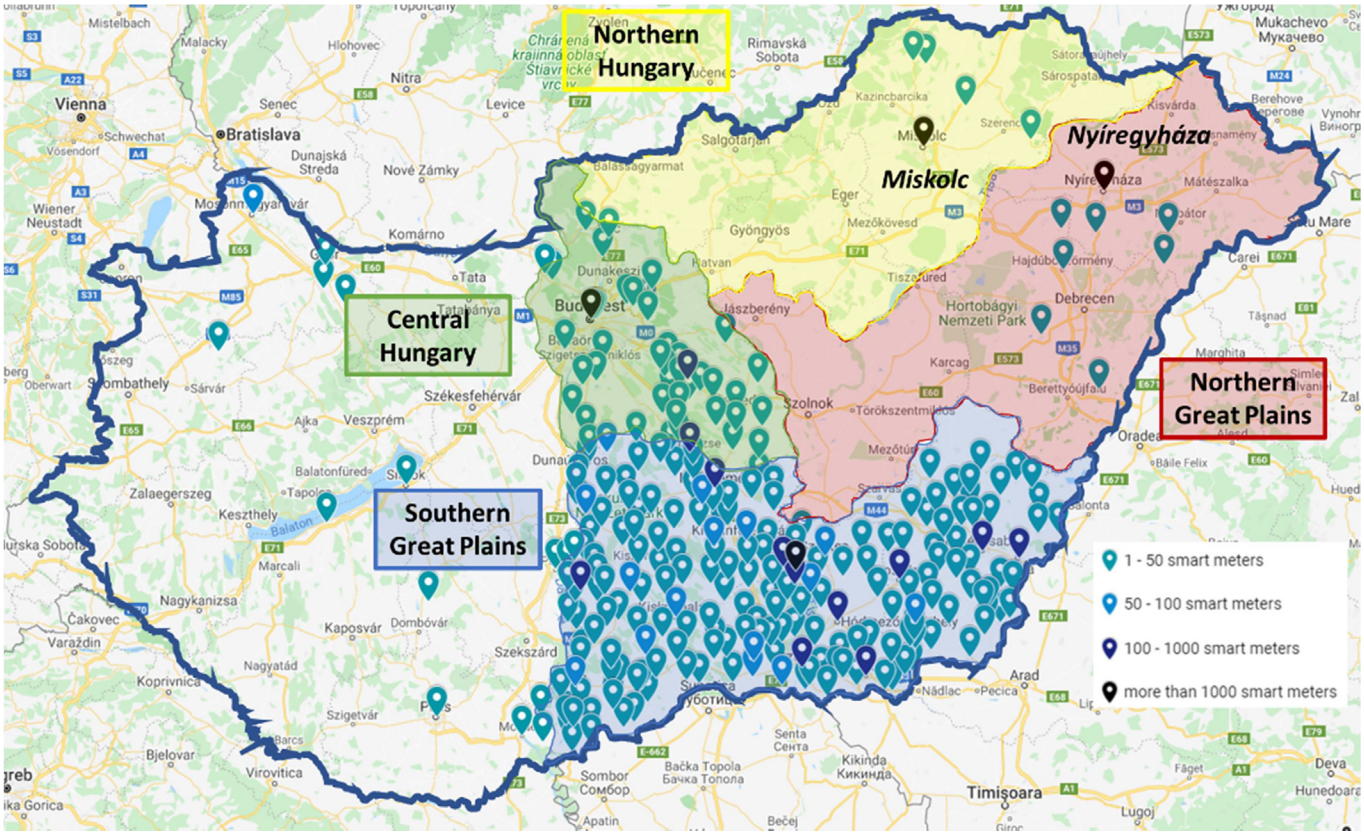


Fig. 1. The geographical distribution of the installed smart meters.

Table 1
Number of meters deployed in the KOM project for the different consumption types.

Consumption type	Total no. meters deployed	Residential meters deployed
Natural gas	22,079	7368
Heat	53,447	53,432
Electricity	*28,993*	*24,917*
Water	24,115	22,231
Σ	128,634	107,948

or unoccupied) and the presence of solar panels/collectors. In this paper, only the information on the size of the building, the building type and condition was used. The three main building categories applied in this paper are single-family homes before and after 1990 and multi-family buildings. Although the survey process involved some manual work, it generally worked well and fast enough. However, in some cases, the building could not be identified, streetview images were missing or blocked by external objects such as trees. In some instances, the house number was unrecognisable. Finally, 1 282 datasets were selected for analysis where both metered data quality was high, and further information on the building was available. Some meters were excluded in a later processing phase due to further irregularities, and 1186 datasets were included in the cluster analysis. For these meters raw data is available from January of 2017 to December of 2018. For some meters, this timespan is shorter, but for all meters included in the investigations the analysed dataset is at least 1 year long. The sampling time for all meters is 15 min.

Out of the 1186 m, 649 were regular meters without photovoltaic grid feed and 370 with photovoltaic feed. 167 m measured only off-peak energy consumption. Out of the 649 regular meters, 158 were paired with an off-peak meter at the same address (Table 3).

The buildings having the 10 lowest and the 10 highest annual electricity consumption were compared separately for single-family and multi-family buildings. In the case of single-family homes, there were significant differences. The houses with shallow energy consumption are usually in poor visible condition (e.g. not renovated old buildings), have smaller floor areas and have bad building envelope characteristics (old windows, uninsulated walls etc.). On the contrary, houses with very high energy consumption usually seem to be recently renovated or newly built. They are bigger buildings and have good building envelope characteristics (insulating windows, insulated walls etc.). The higher number of occupants can explain the higher consumption of renovated/new buildings – e.g. the building has a family in it with at least 4 people – and more household appliances. In old buildings, especially in the countryside, older adults mostly live alone or in couples without children due to the urbanisation effect. In the case of multi-family homes, this kind of conclusions could not be drawn. The flats in the same building have very different electricity demand.

Table 2
Number of electricity meter datasets analysed after filtering and survey.

	Address available	Pre-filtered data	Surveyed address	Surveyed buildings	Analysed meters	Clustered meters
Electricity	*9237*	*4454*	*4090*	*2595*	*1282*	*1186*

Table 3
Types of the analysed electric meters.

Clustering results – analysed electric meters					
1. without photovoltaic grid feed (Group A)	2. off-peak (Group B)	1. + 2.	3. peak meters (off-peak meter also in the unit) (Group C)	4. with photovoltaicgrid feed	1. + 2. + 4.
649	167	816	158	370	1186

In this case, the flats probably have similar properties, the behaviour of the occupants causes the differences.

Regarding the representativeness of the sample analysed in terms of the Hungarian residential building stock, it can be stated that the analysis of 1186 m can provide statistically relevant results with 3% error besides 95% confidence interval [63]. This is considered to be a good representation of the total number of 4 439 959 residential units found at the 2018 local census [64].

In our analyses conducted on settlement types and building types, the error exceeds these levels because, for some categories, the number of meters with appropriate data quality decreased significantly. Only 52% and 54% of meters could be assigned to settlement or building type categories respectively. The error of the statistical representativeness can be seen in Table 4 for the categories mentioned above:

2.1.3. Preliminary data filtering

For the data assessment, we used MATLAB software [65]. Considering the large amount of data, automatic algorithms were developed to facilitate data filtering and processing.

The smart meter data available was filtered first to remove meters with unreliable data (e.g., too many interruptions in collected data, irregular values, or trends). For the filtering process, an algorithm was developed. For this filtering, the meter readings were used, the electricity used from the first data point sampled to the actual point. The most important filtering aspects were as follows:

1. Removal of meters with long interruption periods:
   a. A criterion was established to remove meters with interruptions of longer than 48 h for more than 3% of the data points.
   b. The sampling time calculated was summed for all points. Also, the times lower than 48 h were summed as well. The ratio of these two sums was calculated, and it should be higher than 0.95.
2. Removal of meters with zero or negligible consumption: For filtering out such meters, the mean electric power demand between two measured points was calculated. Meters with electric power equal to zero for more than 99% of the measured points were excluded from further investigations.
3. The number of adjacent days for which at least in one of the points the electric power demand was not zero was calculated. The number of such days had to be higher than 365.

In addition to these steps, the profiles were further filtered during the clustering process.

Table 4
Statistical errors in the categories applied besides 95% confidence interval.

Village	9%	Single home built before 1990	5%
Town	8%	Single home built after 1990	
City	5.5%	Multi-family building	6.5%

2.2. Clustering methods

In this chapter, the used clustering methods are described. In our investigation, three different clustering methods were applied to examine the energy consumption data in MATLAB. The models were built in MATLAB based on our codes, while the existing functions were not used. First, the k-means clustering method was selected because it is commonly used in studies in this field. Second, the fuzzy k-means clustering method was applied, a modified version of the k-means clustering technique. Third, the agglomerative hierarchical clustering technique was used. The hierarchical clustering methods work with a different logic and without iteration, contrary to k-means techniques. Applying the clustering methods, the distance metric for the calculation has to be given. The determination of this metric has a significant impact on the result. For our analysis, the Euclidean distance metric was used based on the literature, which is the most used solution and appropriate for our research. This assumption is reinforced by Li et al [66]. The used distance calculation method was set on the basis of the work of Chicco et al [67]:

$$d(p_1, p_2) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_{n,i} - p_{m,i})^2} \quad (1)$$

where p_n and p_m are the examined profiles, N is the number of samples in the examined time period.

K-means clustering method is a hard clustering technique. It means that every data is ordered into only one cluster. This technique works in iteration steps [68]. First, the number of clusters (k) must be determined, and "k" random initial data has to be chosen as the centroid of the clusters. At the next step, the distance between each data and the cluster centroids must be calculated. Each data is ordered into the cluster from which its distance is minimal. After that, the new cluster centroids must be calculated based on the data belonging to the clusters. The centroid profiles of the clusters were calculated as the mean value of the profiles belongs to the same cluster every 15 min. Then the distances have to be recalculated, and the steps have to be repeated until the given criterion is fulfilled. The results of this method depend on the first random initial cluster centroids. To avoid this problem, basic profiles were chosen based on the consumption data and the agglomerative hierarchical technique, and the results were compared. The final result was chosen based on the cluster validity indexes, described in the next Section. Applying the k-means clustering method, the iteration lasted until no data is ordered into another cluster, but a maximum of 1000 iteration were allowed.

The fuzzy k-means clustering method works similarly to the k-means clustering technique [67]. The difference is that the fuzzy k-means method is a soft clustering technique. It means that every data are ordered not only into one cluster but into every cluster with different probability, given by the membership degree. The membership degree is calculated based on the distances between the data and cluster centroids [60,69]:

$$MD(n,k) = \frac{1}{\sum_{m=1}^{K}\left(\frac{d(p_n,c_k)}{d(p_n,c_m)}\right)^{\frac{2}{\beta-1}}} \quad (2)$$

where p(n) are the consumption profiles, c(m) are the cluster centroids, β is the fuzziness parameter.

The value of the cluster centroids was calculated, taking into account the membership degree. Calculating the membership degree, the fuzziness parameter has to be given, which determines the fuzziness of the clustering method. When higher parameters applied, all data have a stronger influence on the value of all cluster centroids. Even though the recommended and most commonly used value is 2 [69], 1.5 was set in our research as the fuzziness parameter. When the value 2 was used in our investigation, the final cluster centroids were too similar. The clustering became meaningless, and therefore the reduction of this value was necessary. Similarly to the k-means technique, the number of clusters has to be given in the first step. The method results also depend on the first random initial cluster centroids; therefore, the basic profile combinations were selected the same way as in the k-means technique. During the iteration, the maximum distance between cluster centroids was divided by an average value of the cluster centroids for every 15 min of data. An average value of them was determined. The iteration lasted until the changes of the cluster centroids dropped below 0.5%.

The third applied technique is the agglomerative hierarchical clustering method. This method works in steps, and there is no need for iteration [68]. In the first step, every data belong to separate clusters. The distance between each cluster centroids – which are the first step – is calculated, and the two nearest clusters are merged. In this research, the cluster centroids were calculated as the mean value of all data merged into the same cluster. This method could be continued until every profile is merged into one cluster; therefore, the number of clusters does not have to be determined beforehand. Because of its working method, the results of this clustering method are always the same; just the final number of clusters have to be determined manually.

2.3. Clustering results evaluation

Several cluster validity indexes could be used to determine how well a clustering method works on our data and the optimal number of clusters [70]. Three different techniques were applied in our work: the elbow and the silhouette methods and the Dunn index.

The usage of the elbow method is prevalent, even though the number of clusters could not be determined based on objectively calculated values. To apply this technique, the sum or the average of the distances between the cluster centroids and cluster profiles have to be calculated and plotted according to the number of clusters. The "elbow" of this diagram indicates the optimal number of clusters. In our investigation, the sum of distances was calculated and plotted for 1–10 clusters. Due to its empirical features, the determination of the optimal number couldn't have been automated.

To apply the silhouette method, the silhouette score has to be calculated on the basis of distances. The value of this score shows the goodness of clustering, the optimal number of clusters is at the maximum silhouette score. The silhouette score was calculated for every profile. The cohesion (a_i) was determined as the average distance between the examined profile and the other profiles in the same cluster. To determine the separation (b_i), the average distances between the examined profile and the profiles in other clusters were calculated. Therefore *k-1* values were obtained, and the minimum value was chosen to represent the separation. The silhouette score was calculated as in [68]:

$$S_i = \frac{(b_i - a_i)}{\max\{a_i, b_i\}} \quad (3)$$

The silhouette score of the clustering method was determined as the average of the silhouette values of each profile.

To calculate the Dunn index, the optimal number of clusters is indicated by the maximum value. For this analysis, different distance values could be applied [70]. For each clustering result, one Dunn index value was determined in our research. First, the distances between the cluster centroids were calculated, and the minimum value was chosen. Second, the maximal distances between profiles in the same clusters were determined, and the highest value of them was chosen. The Dunn index was calculated as the ratio of the minimal distance between cluster centroids and maximal distance between profiles in the same cluster.

In our previous work, educational building gas consumption data were analysed, and different examination options were compared [71]. Three different measurement profile types were used: normal, simplified and integral. We found the normal profile type as the most appropriate. Therefore, in this research, only this profile type was examined. In our previous paper, k-means, fuzzy k-means and agglomerative hierarchical clustering methods were used as well. The above-mentioned indices were applied to determine the optimal number of clusters: elbow method, silhouette index, and Dunn index. In conclusion, the fuzzy k-means clustering technique proved to be the most suitable method, and the number of clusters was determined the best by the elbow method.

2.4. Workflow and methodology

The methodology of our work is presented in Fig. 2. It shows the different steps of data processing and evaluation.

The meters used for this paper were residential electricity meters. Among this broader group, five subgroups were established for the clustering process according to the sub-metering principle: A) regular electricity meters; B) off-peak meters only; C) regular meters only (in these housing units there is also an off-peak meter); D) regular meters and accompanying off-peak meters summed up (sum of B and C meters); E) merge of group A) and D), which includes all investigated apartments. Further groups could also be defined, but they were excluded from this research (for example, housing units with significant photovoltaic feed to the grid, as apparently photovoltaic panels were installed for these locations).

Group A covers the most widespread case, but the available information is not purpose-specific as it can cover any form of electric use, including electric DHW production (DHW not connected to the off-peak meter in this case) or not. However, this is the most representative dataset among all.

Group B was set up for meters with off-peak controlled electricity. These meters were installed alongside regular electricity meters. Off-peak electricity is used mainly for DHW production in Hungary. These profiles are influenced by the load of the grid and the operational scheme of the grid operator. Group B can give valuable information about the hot water production schedule and off-peak times of the operator.

Results of group C can provide the most reliable information on electricity use without hot water production. It is the most valuable information for building simulation models as profiles for housing appliances and lighting energy demand inputs can be given independently from the hot water demand. Therefore, we focused on this group the most.

Group D is a synthetic group created for investigating the impact of DHW production on the total electricity consumption profile.

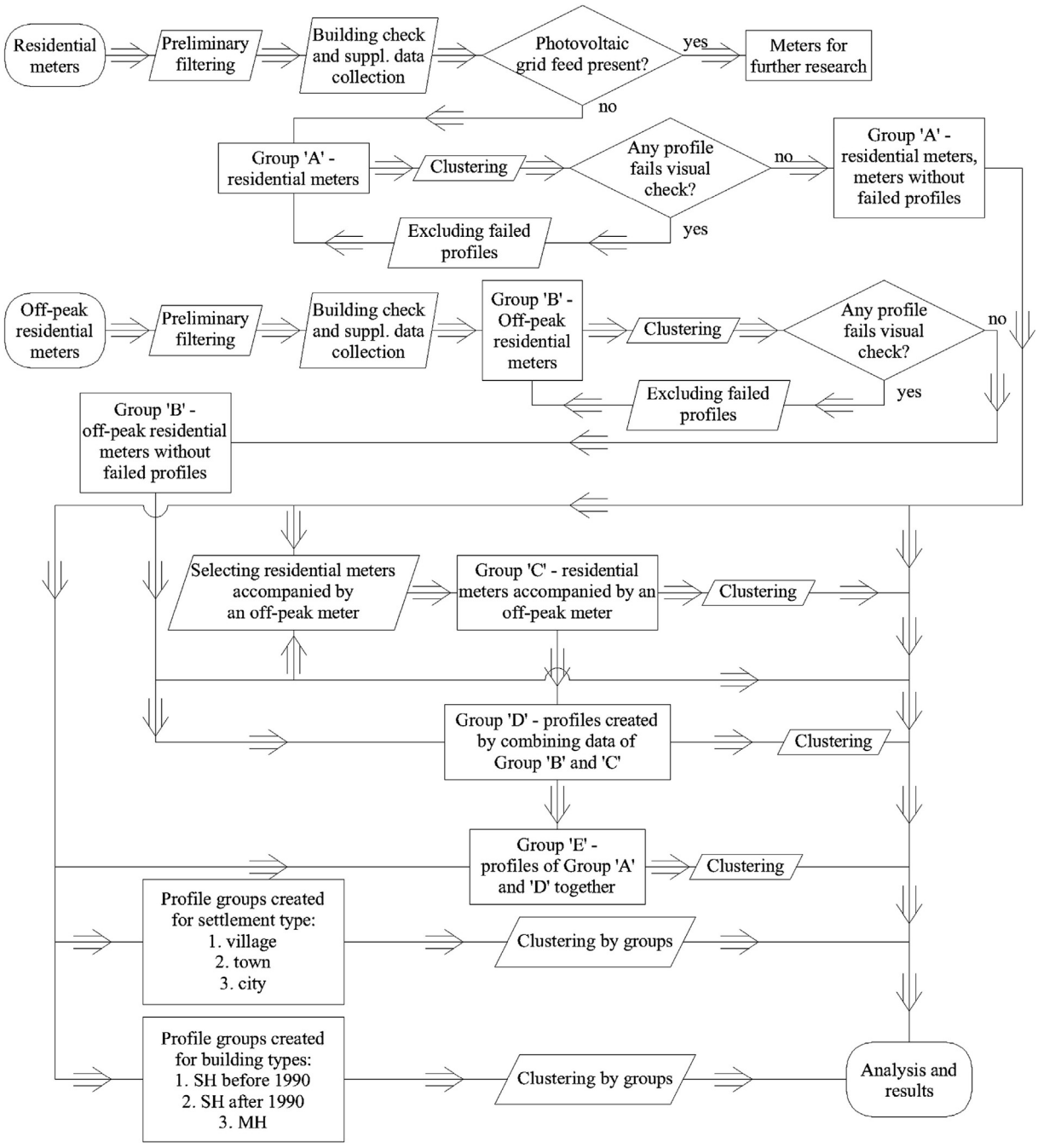


Fig. 2. Workflow for creating the different groups for further analysis of data.

And finally, group E includes the total consumption of all housing units.

In addition to these groups, the regular meters (Group A) were studied in groups based on the location and the building type they were installed in. The number of meters in each group is shown in Fig. 3. In Group E, Group A and Group D are both included.

For each group, "annual profiles" were created. The annual profile represents the monthly mean values of the daily electricity consumptions of the month. It is better than monthly consumptions since it eliminates the different length of the months.

Daily profiles were created from the hourly mean electric power (hourly average of the 15 min samples). For each day, one profile can be determined. As occupancy profiles change during the week, we analysed separately Tuesdays and Wednesdays (representing typical weekdays), Saturdays and Sundays (weekends), and Friday as a transient day before the weekend. Daily profiles can be influenced by the seasons (e.g. cooling energy demand in summer, electric heating in winter); therefore, the daily profiles were checked for other characteristic months (mainly January, April and August, depending on the purpose of the analysis). The coldest month was January, and the hottest one was August in the analysed period. The applied day and month combinations are presented in Table 5.

Our main focus was to determine the time dependence of daily fluctuations rather than absolute consumption values. No information was available about apartments' size and number or composition of occupants. Therefore, specific dimensionless values were calculated for each meter, dividing the hourly values by the annual

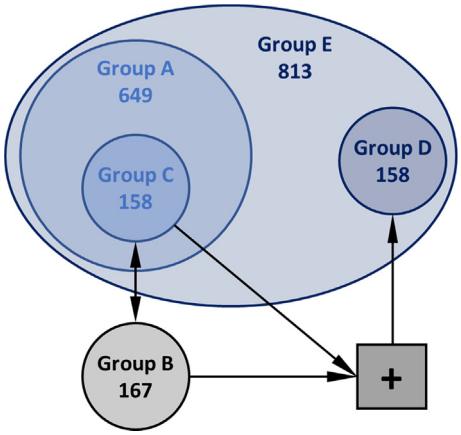


Fig. 3. Number of meters in each group.

Table 5
Days investigated for different months of the year.

Months	All days	Wednesday	Friday	Saturday and Sunday	Tuesday and Wednesday
All year	X	X	X	X	X
January				X	X
April				X	X
June	X				
August	X			X	X
October				X	X

mean value of the corresponding meter. For the annual profile, the monthly values were divided by the annual monthly average.

2.5. Selection between clustering methods

The selection between the different clustering methods can be achieved in two different ways. No uniform recommended solution or cluster validity indexes were found to determine the optimal clustering method in the literature. Therefore, the following two techniques were applied for our examination.

2.5.1. Selection on basis of cluster validity indexes

The cluster assessment metrics, like the total within the sum of distances (for using the elbow method), the silhouette and the Dunn index, can be used for selecting the optimal number of clusters, but for selecting the optimal clustering method as well. Fig. 4 presents the Dunn index metric for clustering, performed on data of group A. The results of the other metrics performed on the same group of data could be found in Annex 1. For this investigation, k-means clustering was performed in two different ways: on the regular way and by applying a constraint on the minimal number of profiles in one cluster; for this case, it was 10. From the figures, it can be seen that regular k-means and hierarchical clustering outperform the modified k-means and fuzzy k-means techniques. However, in the latter two cases, the total sum of distances is much smaller; therefore, from this point of view, they proved to be better methods for this database.

2.5.2. Selection on basis of cluster centroids

The cluster centroids are basically the average profiles of all profiles belonging to the same cluster for the k-means and hierarchical clustering methods. In contrast, in the case of fuzzy k-means clustering, other profiles are used to calculate the centroids. These centroids can be visually assessed, and the clustering method providing the most realistic results can be selected. Examples can be seen in Fig. 5 and Fig. 6 and in Annex 2, which present all daily profiles of the year from group A. The title of the figures indicates the examined time periods: m_i represents the months (1: January, . . ., 12: December), d_j represents the days (1: Sunday, . . ., 7: Saturday). The clusters acquired for the max silhouette were selected for interpreting this selection method. The hierarchical and k-means clustering methods created clusters with a low number of irregular, outlying profiles (Fig. 5). The other two methods produced more representative results in each cluster (Fig. 6). The figures in Annex 2 (Figure A.3 – Figure A.5) show very similar trends:

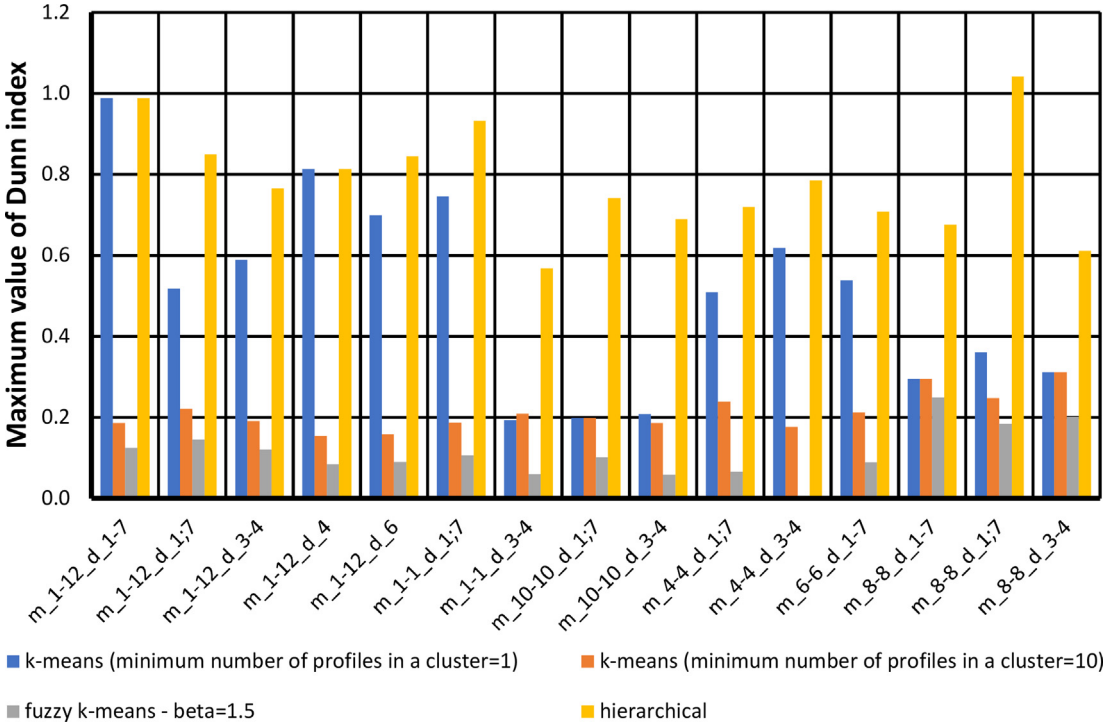


Fig. 4. The maximum value of the Dunn index versus the clustering method for the cases investigated in group A.

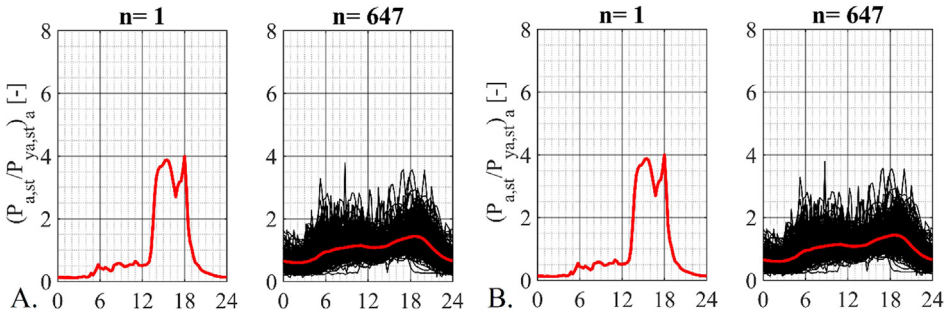


Fig. 5. Cluster centroids and profiles in group A, m_1-12_d_1-7 for two clusters determined with A.: the k-means method with no constraint, B.: the hierarchical method.

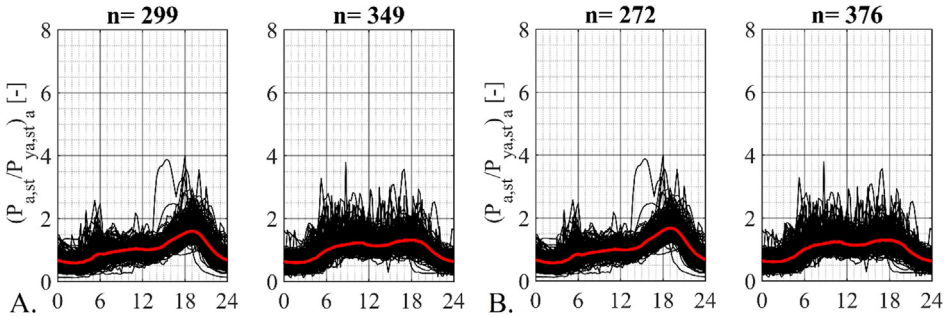


Fig. 6. Cluster centroids and profiles in group A, m_1-12_d_1-7 for two clusters determined with A.: the fuzzy k-means method, B.: the k-means method with constraint of having 10 profiles minimally in a cluster.

while the hierarchical clustering sorts the irregular profiles into small clusters, the modified k-means and fuzzy k-means clustering techniques create larger clusters with similar cluster centroid profiles.

2.5.3. Selecting the number of clusters

Once the clustering method is selected, the number of clusters has to be decided. The optimal number of clusters varies from case to case and has to be decided individually. The following two examples explain how the optimal number of clusters was decided.

The first example shows how the optimal number of clusters was selected for group A, including all daily profiles of the year. As a first step, the total within the sum of distances and the Dunn index and silhouette score values have to be determined for different cluster numbers (Fig. 7). It is worth specifying a maximal number of clusters as a second step because small, non-representative clusters appear above a certain number of clusters. In addition, the high number of clusters is difficult to manage, and its usefulness is questionable. In our example, the optimal number of clusters seemed to be 3 or 4. Results for the two cases are presented in Fig. 8 and Fig. 9. The third step is a visual inspection. Looking at the diagrams, it is easy to conclude that the four clusters case does not provide any additional information than the three clusters case, as obviously two of the four have very similar characteristics. Thus, the three clusters case should be selected in this example.

The second example shows how the optimal number of clusters was selected for group B (off-peak meters), including all daily profiles of the year. In this case, after checking the metrics (Figure A.6) the optimal number of clusters seemed to be 4 or 6 on the basis of the total within the sum of distances. However, for the other two metrics, it was clearly 2, which seemed to be too few after analysing the results manually. Results for the 6 clusters are presented in Figure A.7. By visual inspection, it can be easily recognised that the last cluster contains only meters with zero consumption. Obviously, in these apartments, no appliances are connected to the

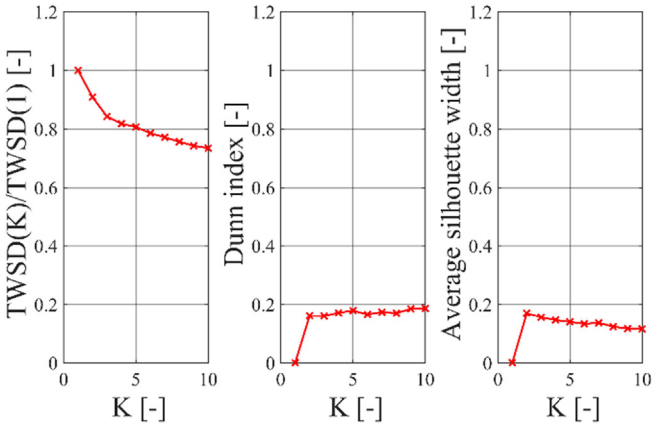


Fig. 7. Clustering metric for k-means method with constraint of having 10 profiles minimally in a cluster for group A, m_1-12_d_1-7.

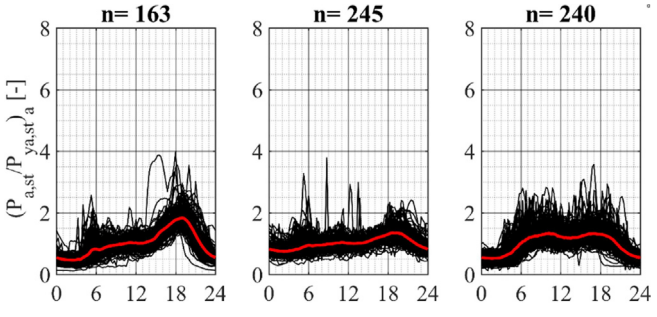


Fig. 8. Cluster centroids and profiles in group A, m_1-12_d_1-7 for three clusters determined with the k-means method with constraint of having minimum 10 profiles in a cluster.

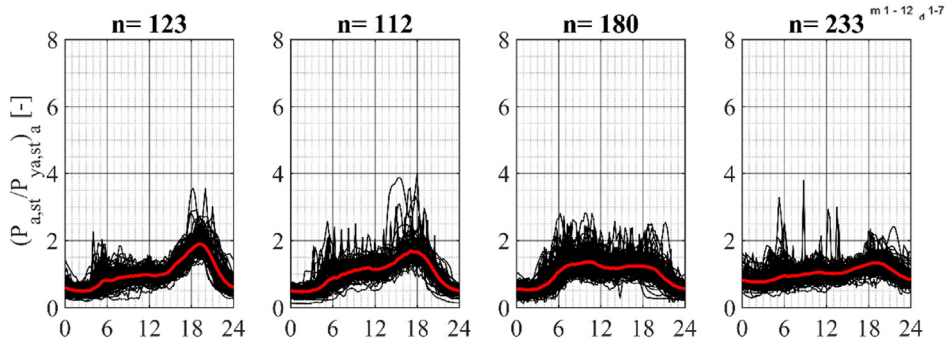


Fig. 9. Cluster centroids and profiles in group A, m_1-12_d_1-7 for four clusters determined with the k-means method with constraint of having minimum 10 profiles in a cluster.

off-peak meter, so they should be removed from the group. After removal, the optimal number of clusters proved to be four, and this clustering result is presented in Figure A.8.

In the visual inspection phase, the most frequent reasons to select the lower number of clusters are: a) too low number of meters are in one cluster, b) two clusters show very similar characteristic or c) irregular meters are in a cluster that should be removed from the population.

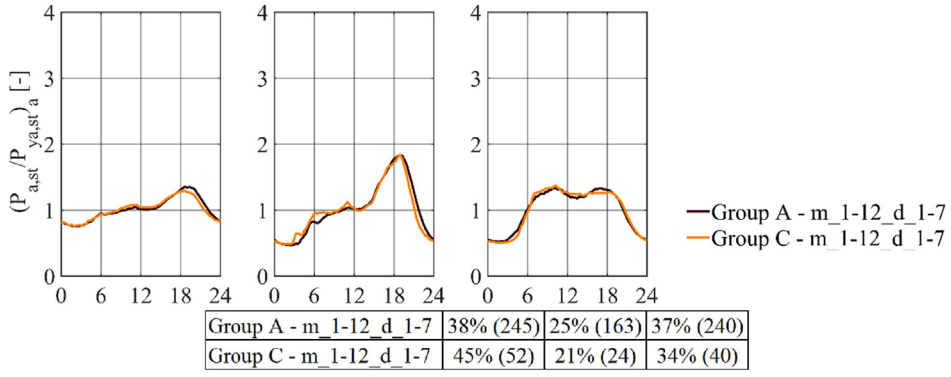
3. Results and discussion

3.1. Analysis of daily profiles

In this chapter, results of daily profiles are presented to identify peak and low demand periods during the day and determine typical occupant profiles. Our goal was to investigate the impact of the following parameters on energy profiles: group class (A, B, C, D, E), day of the week (mainly weekdays/weekends), season, settlement type and building type. Average profiles were determined for each group and parameter to illustrate the daily evolution of electric energy consumption. All diagrams are dimensionless: in the daily diagrams, the measured values are divided by the annual daily average consumption; in the annual diagrams, they are divided by the annual monthly average value. Thus, the detection of fluctuations and amplitudes within the period is not disturbed by the differences between consumption magnitudes.

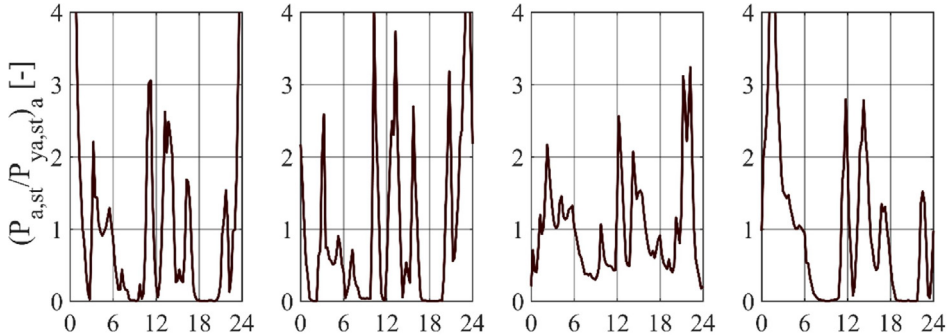
3.1.1. Comparison of results per group

Results per group are presented in Fig. 10-Fig. 12. The tables under the figures show the number of profiles used to create the



Group A - m_1-12_d_1-7	38% (245)	25% (163)	37% (240)
Group C - m_1-12_d_1-7	45% (52)	21% (24)	34% (40)

Fig. 10. Cluster centroids and profiles in group A and C, m_1-12_d_1-7 for three clusters determined with the k-means method with constraint of having 10 and 5 profiles minimally in a cluster, respectively.



Group B - m_1-12_d_1-7	22% (28)	14% (18)	35% (44)	28% (35)

Fig. 11. Cluster centroids and profiles in group B, m_1-12_d_1-7 for four clusters determined with the k-means method with constraint of having 1 profiles minimally in a cluster.

cluster centroids presented in the figures. Group A and Group C show up nearly identical performance (Fig. 10), meaning that those apartments, which are not equipped with an off-peak meter, use similar electric appliances as those in Group C. Therefore, it means that apartments without off-peak meters are mostly not using electricity for DHW production. Otherwise, their profile should be different. Out of the three clusters, two show similar trends with an evening peak. The difference is the magnitude of the peak; evening consumption is significantly higher for one cluster, which might be caused by more people arriving home after office hours and also which was found to be a characteristic of one-person households and two-bedroom apartments by our earlier review [72]. The third cluster shows a more balanced consumption during the day, which is likely to be caused by users staying at home all day.

Fig. 11 presents clustering results for off-peak meters only (Group B). Here, consumption is not determined only by the demand but by the supply as DHW tanks are heated when the utility company provider feeds electricity into the off-peak circuit. In fact, according to the graphs, in terms of the shape of the profiles, the feed is the determinant, not the demand. Thus, it is not possible to assign characteristic consumer behaviour to different clusters. The shift between load and demand is due to the tanks. In order to determine the demand profile for DHW, it would be more beneficial to analyse water consumption trends in future research.

Taking a look at Fig. 12, we can conclude that the impact of DHW production connected to the off-peak circuit is dominant on the total consumption. Consequently, the problem described in Figure B is inherited by groups D and E. Therefore, the power supply of the utility company takes the dominant impact on consumption, and the profile of the demand cannot be clearly separated due to the time shifts caused by the buffer tank. It would be more appropriate to determine the DHW demand profile based on the water consumption data, which is out of the current work scope.

It can be concluded that the analysis of group C is the most expedient for determining the daily electrical consumer demand profiles. Nearly the same results can be expected from the analysis of Group A, which has a higher sample number, therefore in the next sub-chapters, we focused on the results of Group A. However, as we will see later, the situation is different when examining annual profiles.

3.1.2. Day of the week

As occupancy is significantly different in many residential buildings between weekdays and weekends, we performed separate clustering analysis for weekends (Saturdays and Sundays, Fig. 13) and weekdays (Tuesdays and Wednesdays, Fig. 14). The weekday analysis was restricted to two days so that the same sample number could be achieved for the two cases. Furthermore, the impact of particularities of the start and the end of the working period (Monday and Friday) could be eliminated in such a way.

Fig. 15 merges results for both cases and the "all days of the week" as well. To find the different groups' similar profiles, we compared them one by one by calculating the minimum distance to the average profile using iteration. Additionally, the number of buildings belonging to the cluster centroid is shown in the table below the graph.

Interestingly, there is only a moderate difference between the shapes of the weekdays (Wednesday-Thursday) and weekend

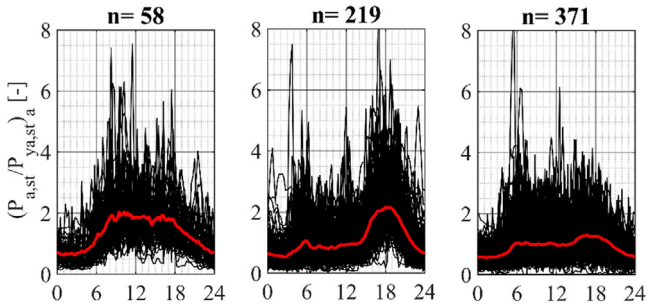


Fig. 13. Cluster centroids and profiles in group A, m_1-1_d_1;7 for three clusters determined with the k-means method with constraint of having 10 profiles minimally in a cluster.

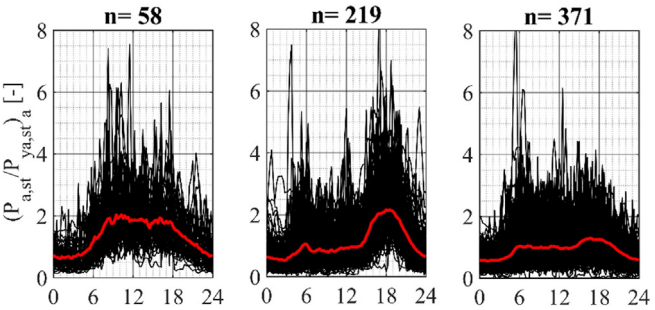
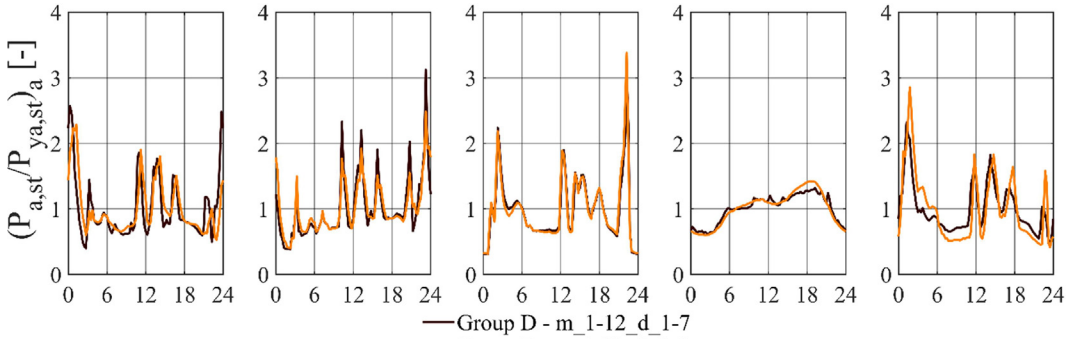
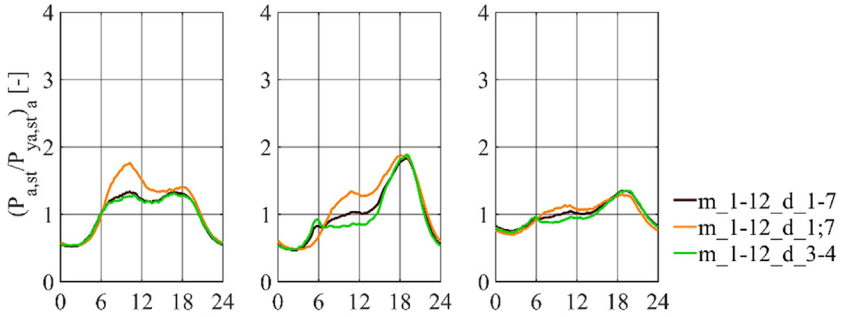


Fig. 14. Cluster centroids and profiles in group A, m_1-1_d_3-4 for three clusters determined with the k-means method with constraint of having 10 profiles minimally in a cluster.



Group D - m_1-12_d_1-7	11% (18)	9% (15)	9% (14)	54% (85)	16% (26)
Group E - m_1-12_d_1-7	4% (29)	3% (24)	2% (18)	90% (728)	2% (13)

Fig. 12. Cluster centroids and profiles in group D and E, m_1-12_d_1-7 for five clusters determined with the k-means method with constraint of having 10 and 5 profiles minimally in a cluster, respectively.



m_1-12_d_1-7	37% (240)	25% (163)	38% (245)
m_1-12_d_1;7	34% (223)	19% (124)	46% (301)
m_1-12_d_3-4	35% (229)	27% (177)	37% (242)

Fig. 15. Cluster centroids compared in group A for m_1-12 with numbers of profiles in each cluster, determined with the k-means method with constraint of having 10 profiles minimally in a cluster.

(Saturday-Sunday) profiles. Still, there is a significant shift in the number of dwellings in each cluster. From this, we can conclude that some consumers are switching to another profile when the weekend comes. Between Wednesdays and Fridays, no remarkable difference could be found (Fig. 16).

Results for the off-peak profiles are presented in Annex 4, and one can see even less difference between weekdays and weekends.

3.1.3. Seasonal

To examine the effect of seasonality on the daily profile, a separate cluster analysis was performed for the average days of January, April, August, and October, where August was the warmest period and January was the coldest period. The results for the weekends are shown in Fig. 17 (including the annual mean profiles as well). The peaks in January and August are slightly higher than in the other two months, but this is not significant. However, there is an outstanding cluster with exceptionally high consumption in August, which includes 52 flats. Indeed, air conditioning systems were used in these apartments, the effect of which is therefore clearly visible.

Further results can be found in Annex 5. It can be stated that the weekday results are similar to the weekend ones.

3.1.4. Geographical location, settlement types

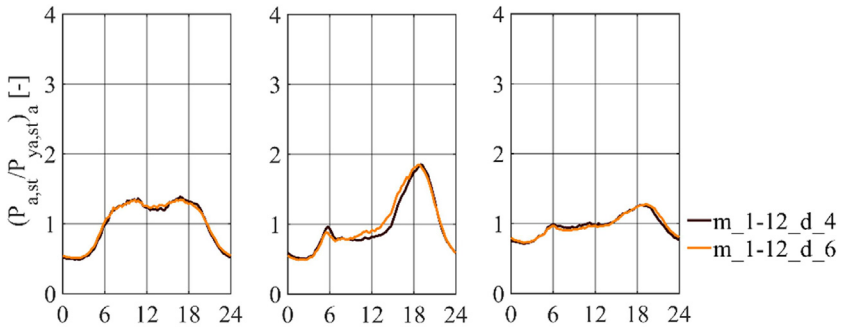
We investigated if the settlement type influences the user profile. Three settlement types were analysed: villages, towns and cities. We used the Hungarian Central Statistical Office database to check the type of settlement [73]. Fig. 18 shows the yearly profiles for the weekends divided into three clusters. The profiles are somewhat similar except for villages in the second cluster (5 housing units), which shows an afternoon peak occurring earlier than in towns and cities. In the first cluster, the relative morning peak is the highest in villages, followed by towns and cities. However, it should be noted that the number of meters in the first cluster is very low for villages. In general, the consumption data do not support the belief that people would rise significantly earlier in villages to towns and cities.

Fig. 19 shows two weekdays, Tuesday and Wednesday. The second and third clusters show similar profiles, but in the first cluster, significant differences can be observed when the peak time is different for each settlement type.

Fig. 20 shows the cluster centroids for all days of the year. Here no significant differences can be observed between the settlement types which conflicts the findings of our previous review [72].

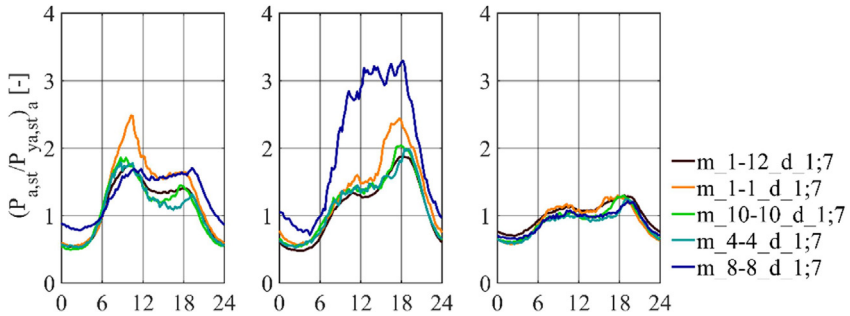
3.1.5. Building type

Different building types might result in different user behaviour, or different user types might choose different buildings to live in. Therefore, the impact of building type was investigated for three subgroups: old (built before 1990) and new (built in 1990 or later) single-family homes and multi-family dwellings. The creation of additional subgroups was discarded because the number of meters in at least one subgroup would have been too small.



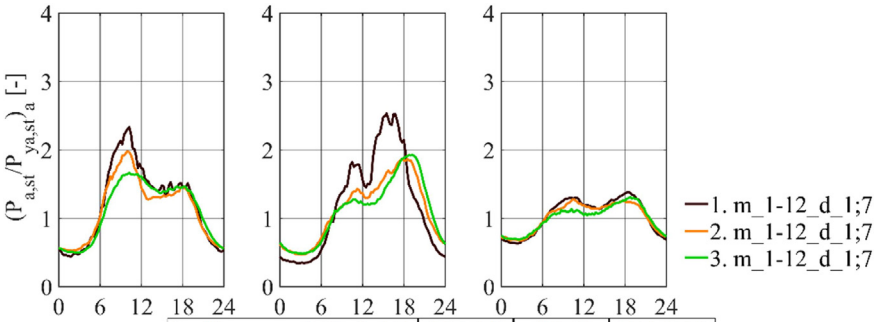
m_1-12_d_4	27% (173)	31% (199)	43% (276)
m_1-12_d_6	35% (225)	24% (155)	41% (268)

Fig. 16. Cluster centroids compared for Wednesdays and Fridays (in group A for m_1-12 with numbers) of profiles in each cluster, determined with the k-means method with constraint of having 10 profiles minimally in a cluster.



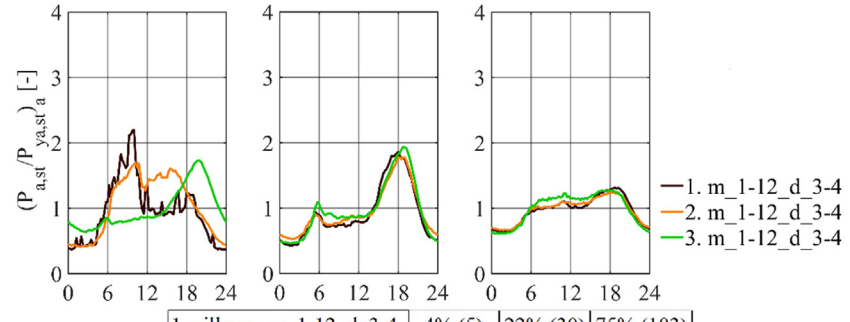
m_1-12_d_1;7	34% (223)	19% (124)	46% (301)
m_1-1_d_1;7	26% (169)	19% (126)	54% (353)
m_10-10_d_1;7	24% (155)	21% (138)	55% (355)
m_4-4_d_1;7	25% (162)	19% (126)	56% (360)
m_8-8_d_1;7	44% (286)	8% (52)	48% (310)

Fig. 17. Cluster centroids compared in group A for d_1;7 with numbers of profiles in each cluster, determined with the k-means method with constraint of having 10 profiles minimally in a cluster.



1. village - m_1-12_d_1;7	17% (23)	4% (5)	80% (110)
2. town - m_1-12_d_1;7	26% (45)	16% (29)	58% (102)
3. city - m_1-12_d_1;7	26% (88)	22% (74)	51% (172)

Fig. 18. Cluster centroids compared in group A for different settlement types for m_1-12_d_1;7 with numbers of profiles in each cluster, determined with the k-means method with constraint of having 5 profiles minimally in a cluster.

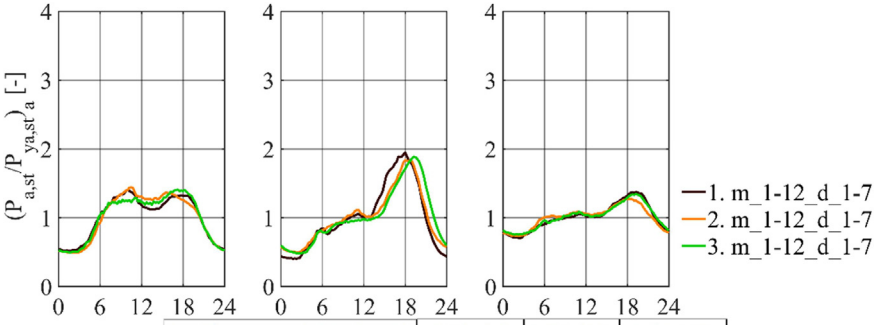


1. village - m_1-12_d_3-4	4% (5)	22% (30)	75% (103)
2. town - m_1-12_d_3-4	9% (16)	30% (53)	61% (107)
3. city - m_1-12_d_3-4	29% (97)	23% (77)	48% (160)

Fig. 19. Cluster centroids compared in group A for different settlement types for m_1-12_d_3-4 with numbers of profiles in each cluster, determined with the k-means method with constraint of having 5 profiles minimally in a cluster.

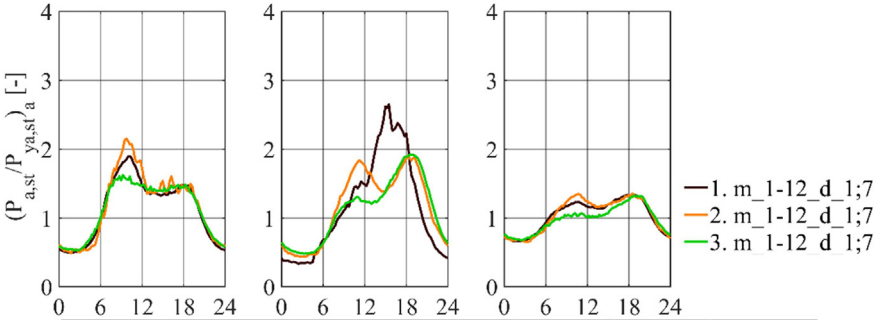
Fig. 21 and Fig. 22 show the profiles for these subgroups of residential buildings divided into three clusters for weekends and all week. The profiles in each subfigure were grouped so that the smallest difference occurs between them using the same distance metric as in the clustering process. Apparently, the profiles were almost the same for the weekends, except the second profile for the single-homes before 1990 for weekends, which characterises only 8 buildings.

Fig. 22 shows the cluster centroids for the resulting profiles of all days of the year. On this basis, there is no difference between the different building types.



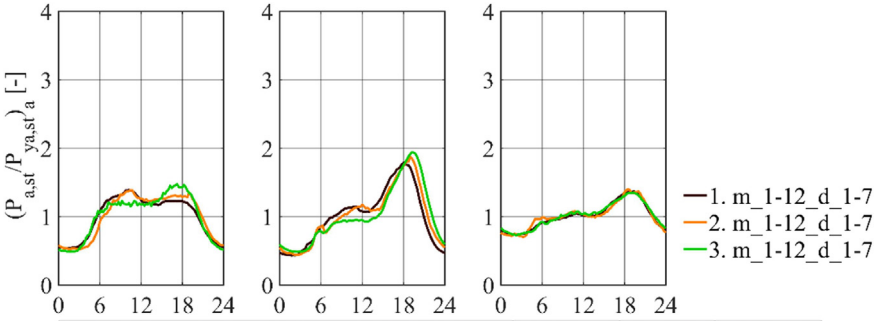
1. village - m_1-12_d_1-7	41% (56)	14% (19)	46% (63)
2. town - m_1-12_d_1-7	30% (52)	23% (41)	47% (83)
3. city - m_1-12_d_1-7	34% (112)	28% (93)	39% (129)

Fig. 20. Cluster centroids compared in group A for different settlement types for m_1-12_d_1-7 with numbers of profiles in each cluster, determined with the k-means method with constraint of having 5 profiles minimally in a cluster.



1. single home before 1990 - m_1-12_d_1;7	30% (98)	2% (8)	67% (216)
2. single home in 1990 or later - m_1-12_d_1;7	19% (13)	20% (14)	61% (42)
3. multi home - m_1-12_d_1;7	26% (62)	25% (59)	48% (113)

Fig. 21. Cluster centroids compared in group A for different building types for m_1-12_d_1;7 with numbers of profiles in each cluster, determined with the k-means method with constraint of having 5 profiles minimally in a cluster.



1. single home before 1990 - m_1-12_d_1-7	34% (111)	23% (74)	43% (137)
2. single home in 1990 or later - m_1-12_d_1-7	36% (25)	22% (15)	42% (29)
3. multi home - m_1-12_d_1-7	30% (71)	27% (64)	42% (99)

Fig. 22. Cluster centroids compared in group A for different building types for m_1-12_d_1-7 with numbers of profiles in each cluster, determined with the k-means method with constraint of having 5 profiles minimally in a cluster.

3.1.6. Generalisability of results

As discussed in chapter 2.1.2 the dataset can provide statistically relevant results with 3% error besides 95% confidence interval for Hungarian building stock. As we could see here in chapter 3, the settlement type and building type does not have a significant influence on the results. Therefore we can assume that results might be generalised for other countries and regions, although a validity check was not possible within the project due to capacity limits. The hypothesis could be justified if cross-checks were performed on data from other countries according to the methodology outlined in the article.

3.2. Analysis of annual profiles

3.2.1. Group a

User demands change throughout the year, for example, due to the use of air conditioners in summer and possibly electric heating in winter. Occupancy rates also show seasonal changes due to holidays and school breaks. The annual profile reflects in which months higher and lower consumption occurs.

Fig. 23 and Fig. 24 show the clustering results for group A, applying the modified k-means clustering technique. The figures represent the annual electricity consumption profiles. Three typical clusters and, therefore, three typical usage patterns could be separated. The users belonging to the first cluster in Fig. 23 are supposed to have air conditioning systems to reduce the heat loads during the summer period. The energy consumption of this equipment could increase the electricity demand in summer. The users belonging to the second cluster on the same figure are supposed to have electric heaters. The usage of them could increase the energy demand during the colder months essentially. The users belonging to the third cluster are supposed to have no electric air cooling or heating equipment or they are applied only occasionally.

Fig. 24 contains the annual cluster centroids for different time periods: black lines indicate the whole week, orange lines represent the weekends (Saturday, Sunday), green lines represent the weekdays (Tuesday, Wednesday). Therefore, the differences caused by the different occupancy could be analysed. Only a slight deviation could be observed between the cluster centroids in the case of second and third clusters. In the case of the first cluster, the profiles are slightly different: during the weekends, a smaller amplitude occurs.

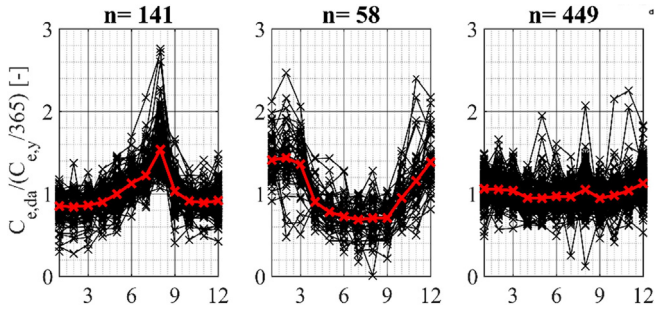
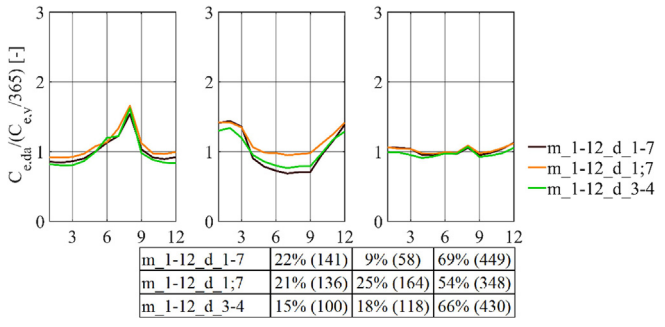


Fig. 23. Monthly cluster centroids and profiles in group A, m_1-12_d_1-7 for three clusters determined with the k-means method with constraint of having 10 profiles minimally in a cluster.



m_1-12_d_1-7	22% (141)	9% (58)	69% (449)
m_1-12_d_1;7	21% (136)	25% (164)	54% (348)
m_1-12_d_3-4	15% (100)	18% (118)	66% (430)

Fig. 24. Monthly cluster centroids compared in group A for m_1-12 with numbers of profiles in each cluster, determined with the k-means method with constraint of having 10 profiles minimally in a cluster.

3.2.2. Group B

In the figures in Annex 6, the clustering results of group B annual data can be seen when k-means clustering was applied. It is evident that the second cluster contains the meters, which are connected to devices with no or minimal electricity consumption. Removing these profiles from the database, the clustering could give more detailed results (Figure A.14). In general, the off-peak consumption of these buildings increases in colder months. It could be related to the fact that these meters measure mainly the electricity demand of DHW devices and that the DHW consumption is larger in winter. These tendencies can be seen in Fig. 25. Monthly average daily consumption compared to the daily average calculated from annual consumption based on the cluster centroid belonging to the highest number of profiles in Figure A.14 from Annex 6.

3.3. Typical profiles

One of our objectives was to provide input data for dynamic building simulation software for a typical daily electrical profile. The simulation calculates the heating and cooling demand based on the building's physical model and the building services systems, and the modelled equipment significantly influences the DHW consumption (particularly buffering). Then, we considered it appropriate to use a profile that is independent of these needs and purely reflects the other consumption (mainly household appliances and lighting). The April data of group C is the most suitable for this purpose.

For simulation purposes, we only present results that reflect all days of the week.The average profile is not suitable for a typical profile because averaging attenuates the amplitudes of the fluctuations. We used two methods to determine the simulation profile; both have advantages and disadvantages.

In the first method (Method A, see Fig. 26), we aimed to ensure that the result faithfully reflected daily maximal values and the daily averages. The former is important for sizing; the latter is essential for modelling energy consumption. To do this, we shifted the average profile to the x-axis, magnified the values so that the maximal value coincided with the average of the maximal values, and then shifted the curve back to the mean value. The disadvantage of this method is that the minimal values thus do not reflect the average minimal values. A negative value also occurred in a few cases, in which case the negative values were considered zero. Thus, the mean value deviated slightly from the original mean; the resulting error is not significant (4% error in the case of the third cluster of Fig. 26, no error for the other two clusters) and is indicated in the caption.

In the second case (Method B, see Fig. 27), we aimed to specify a 90 % probability band in which the vast majority of the occurring values fall. Thus, a minimal and a maximal curve has been determined. Depending on the purpose of the simulation, it can be

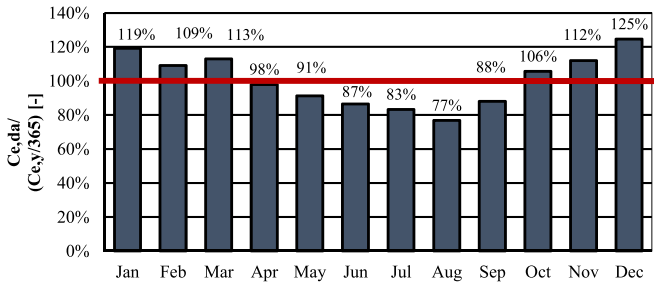


Fig. 25. Monthly average daily consumption compared to the daily average calculated from annual consumption.

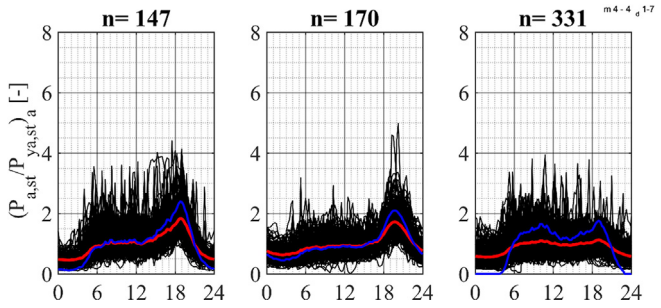


Fig. 26. Simulation profiles using Method A (group A for three cluster for m_4-4_d_1-7, month April).

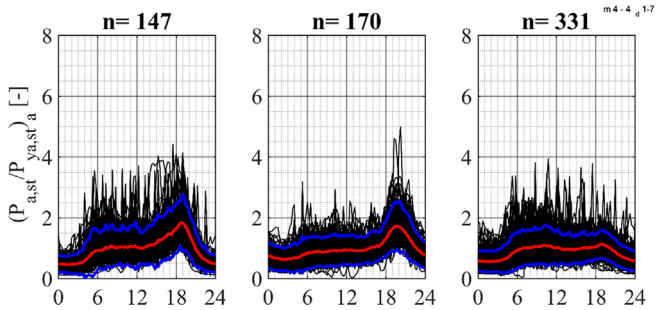


Fig. 27. Simulation profiles using Method B (group A for three cluster for m_4-4_d_1-7, month April).

decided whether it is appropriate to use the minimal or maximal curve. For example, if we intend to use electricity consumption to affect summer internal heat loads, choosing the maximal value guarantees a mistake favouring safety. The same is recommended for safe modelling of total energy consumption. However, to account for winter heat gains for modelling heating demand, the minimal curve is the safe choice. In this case, a statistical approach was applied. This method was developed by assuming that data of profiles belonging to the same cluster can be used to construct the Student's t distribution for each time sampled. This means that for each time sampled, the mean value (which is the value of the cluster centroid for the time) and the standard deviation is calculated. For each time, using the inverse cumulative distribution function, the percentile belonging to the 5% and 95% probability is calculated. If the data is of the Student t distribution, 90% of the data is within these two values, so a higher and a lower bound is constructed for the cluster centroid.

4. Conclusions

In the paper, electricity consumption data of single-family houses and flats were analysed, and consumer profiles were determined by applying clustering techniques. Although we had much more data sets available, we narrowed our findings to 816 residential units because building type information and high-quality datasets were available for those units only. Daily and annual energy consumption profiles were determined; thus, different consumer groups could be distinguished.

In the investigation, MATLAB software was used to analyse the energy consumption of residential buildings. Three different clustering methods (k-means, fuzzy k-means, agglomerative hierarchical) and three different cluster validity indices (elbow method, silhouette method, Dunn index) were applied. To determine the similarity of the energy profiles, the Euclidean distance metric was used. The optimal clustering method and the optimal number of clusters were determined based on cluster validity indices and the shape of the cluster centroids. The best clustering method for our examination proved to be the k-means clustering technique. Analysing the annual and daily consumption data, the optimal number of the clusters was 3 in most cases.

We examined the effect of specific parameters on profiles, such as meter type (regular, off-peak), day of the week, seasonality, settlement type or building type. As little information was available about buildings outside the data sets, this limited the range of parameters that could be examined. The main findings are as follows:

- Concerning the daily profiles, three types of definite profiles can be distinguished, which can be justified by the different occupancy schedules and behavioural habits. One of these can be characterised by a more even consumption throughout the day; the two others had definitive peaks in the morning or/and in the evening. However, the shape of the latter two profiles did not develop in the same way in all cases.
- In the case of off-peak meters, it was impossible to explain the differences between the profiles by demand-side drivers. Instead, it can be read from the profiles during which periods the service providers intervene to supply electricity. This is helpful information to consider when modelling DHW systems with storage.
- In terms of seasonality, the summer-day profiles clearly separated the units using mechanical cooling. Still, where there was no mechanical cooling, the profiles showed a similar course as in the rest of the year.
- There was only a moderate difference between the types of settlements (village, town, city).
- Similarly, there was only a slight difference between the profiles of condominiums, old single-family houses and new single-family houses.

In the annual analysis, three distinct profiles could be distinguished as well. A more balanced consumption can characterise one with a summer peak (presumably due to mechanical cooling), one with a winter peak (presumably electric heating or somewhat heating assistance).

Based on the annual off-peak consumption profile, the hot water consumption is lower than average in summer and higher than average in winter. The average can be best characterised by consumption in April, followed by October. The determined numerical data can be well used for the monthly distribution of consumption when only annual data is available.

Finally, we proposed electrical profiles for dynamic simulation after finding that the April data were most suitable for this purpose. There have been fours profile types determined, which can be used for building energy demand simulation, summer heat load and winter heating demand calculations. The profiles are presented in Figure A.14, and the values can be found in Table A. 1. The yearly profile variability factors are shown in Fig. 24, and the values are given in Table A. 2.

There are several opportunities to continue the research, of which the following would be highlighted:

- Our examination was narrowed to the meters where the type of building was known in the current research. This information was not always necessary during the parameter analysis, so some of the results of the research can be extended to about 4000–5000 housing units, which would strengthen the representativeness of some of our findings.
- Questionnaire surveys could clarify the reasons influencing the development of individual profiles; however, questionnaire surveys are greatly hampered by the GDPR requirements.

- The calculation model could be applied for other countries and regions even for smaller datasets to see whether electric consumption habits are similar or different. Our hypothesis is that the achieved results can be generalised for other countries as well, at least in the region, because not significant difference could be detected between different building and settlement types."
- The profile analysis can be implemented for residential units with solar meters in the future, for which we have a large number of measurement data.
- We plan similar research on gas consumption, water consumption and heat consumption data and extend the investigation to non-residential buildings.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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 by 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.

Adrián Mota-Babiloni acknowledges the financial support of the Valencian Government by the postdoctoral contract APOSTD/2020/032 and of EIT Climate-KIC through the "Pioneers into Practice 2019" programme.

This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enbuild.2021.111376.

References

[1] A. Mahdavi, C. Pröglhöf, User behaviour and energy performance in buildings, 6, Int. Energiewirtschaftstagung an Der TU Wien. (2009) 1–13.
[2] IEA EBC - Annex66, Occupants Behavior Research Bibliography, (n.d.). http://annex66.org/?q=biblio.
[3] W. O'Brien, A. Wagner, M. Schweiker, A. Mahdavi, J. Day, M.B. Kjærgaard, S. Carlucci, B. Dong, F. Tahmasebi, D. Yan, T. Hong, H.B. Gunay, Z. Nagy, C. Miller, C. Berger, Introducing IEA EBC annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation, Build. Environ. 178 (2020), https://doi.org/10.1016/j.buildenv.2020.106738 106738.
[4] M.B. Kjærgaard, O. Ardakanian, S. Carlucci, B. Dong, S.K. Firth, N. Gao, G.M. Huebner, A. Mahdavi, M.S. Rahaman, F.D. Salim, F.C. Sangogboye, J.H. Schwee, D. Wolosiuk, Y. Zhu, Current practices and infrastructure for open data based research on occupant-centric design and operation of buildings, Build. Environ. 177 (2020), https://doi.org/10.1016/j.buildenv.2020.106848 106848.
[5] I.P. Panapakidis, T.A. Papadopoulos, G.C. Christoforidis, G.K. Papagiannis, Pattern recognition algorithms for electricity load curve analysis of buildings, Energy Build. 73 (2014) 137–145, https://doi.org/10.1016/j.enbuild.2014.01.002.
[6] M.S. Piscitelli, S. Brandi, A. Capozzoli, Recognition and classification of typical load profiles in buildings with non-intrusive learning approach, Appl. Energy. 255 (2019), https://doi.org/10.1016/j.apenergy.2019.113727 113727.
[7] M. Bourdeau, P. Basset, S. Beauchêne, D. Da Silva, T. Guiot, D. Werner, E. Nefzaoui, Classification of daily electric load profiles of non-residential buildings, Energy Build. 233 (2021), https://doi.org/10.1016/j.enbuild.2020.110670 110670.
[8] Directive (EU) 2018/2001 of the European Parliament and of the Council on the promotion of the use of energy from renewable sources (recast), Off. J. Eur. Union. (2018) 82–209.
[9] DIRECTIVE 2009/28/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC, Off. J. Eur. Union. 140 (2009) 16–62.
[10] DIRECTIVE (EU) 2018/844 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency, Off. J. Eur. Union. 156 (2018) 75–91. 10.1007/3-540-47891-4_10.
[11] DIRECTIVE 2010/31/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 May 2010 on the energy performance of buildings (recast), Off. J. Eur. Union. 153 (2010) 13–35.
[12] P. Van Aubel, E. Poll, Smart metering in the Netherlands: What, how, and why, Int. J. Electr. Power Energy Syst. 109 (2019) 719–725, https://doi.org/10.1016/j.ijepes.2019.01.001.
[13] C. Brooks, Smarter metering, Renew. Energy Focus. 15 (5) (2014) 16–19, https://doi.org/10.1016/S1755-0084(14)70114-0.
[14] REPORT FROM THE COMMISSION Benchmarking smart metering deployment in the EU-27 with a focus on electricity, 2014.
[15] European Union, Directive of 2009/72/EC of the European Parliament and of the Council of 13 July 2009 Concerning Common Rules for the Internal Market in Electricity and Repealing Directive 2003/54/EC, Off. J. Eur. Union. 211 (2009) 55–93.
[16] ACER/CEER - Annual Report on the Results of Monitoring the Internal Electricity and Natural Gas Markets in 2017 - Consumer Empowerment Volume, 2018.
[17] P. Carvalho, Smart metering deployment in Brazil, Energy Procedia. 83 (2015) 360–369, https://doi.org/10.1016/j.egypro.2015.12.211.
[18] S. Zhou, M.A. Brown, Smart meter deployment in Europe: A comparative case study on the impacts of national policy schemes, J. Clean. Prod. 144 (2017) 22–32, https://doi.org/10.1016/j.jclepro.2016.12.031.
[19] S. Hielscher, P. Kivimaa, Governance through expectations: Examining the long-term policy relevance of smart meters in the United Kingdom, Futures 109 (2019) 153–169, https://doi.org/10.1016/j.futures.2018.06.016.
[20] Y. Kabalci, A survey on smart metering and smart grid communication, Renew. Sustain. Energy Rev. 57 (2016) 302–318, https://doi.org/10.1016/j.rser.2015.12.114.
[21] R. Razavi, A. Gharipour, M. Fleury, I.J. Akpan, Occupancy detection of residential buildings using smart meter data: A large-scale study, Energy Build. 183 (2019) 195–208, https://doi.org/10.1016/j.enbuild.2018.11.025.
[22] Y. Kiguchi, Y. Heo, M. Weeks, R. Choudhary, Predicting intra-day load profiles under time-of-use tariffs using smart meter data, Energy 173 (2019) 959–970, https://doi.org/10.1016/j.energy.2019.01.037.
[23] 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. 217 (2019) 702–715, https://doi.org/10.1016/j.jclepro.2019.01.229.
[24] C. Roach, Estimating electricity impact profiles for building characteristics using smart meter data and mixed models, Energy Build. 211 (2020), https://doi.org/10.1016/j.enbuild.2019.109686 109686.
[25] H. Li, Z. Wang, T. Hong, A. Parker, M. Neukomm, Characterizing patterns and variability of building electric load profiles in time and frequency domains, Appl. Energy. 291 (2021), https://doi.org/10.1016/j.apenergy.2021.116721 116721.
[26] B. Najafi, M. Depalo, F. Rinaldi, R. Arghandeh, Building characterization through smart meter data analytics: Determination of the most influential temporal and importance-in-prediction based features, Energy Build. 234 (2021), https://doi.org/10.1016/j.enbuild.2020.110671 110671.
[27] J. Zhu, Y. Shen, Z. Song, D. Zhou, Z. Zhang, A. Kusiak, Data-driven building load profiling and energy management, Sustain. Cities Soc. 49 (2019), https://doi.org/10.1016/j.scs.2019.101587 101587.
[28] H. Burak Gunay, Z. Shi, I. Wilton, J. Bursill, Disaggregation of commercial building end-uses with automation system data, Energy Build. 223 (2020) 110222. 10.1016/j.enbuild.2020.110222.
[29] M. Samadi, J. Fattahi, Energy use intensity disaggregation in institutional buildings – A data analytics approach, Energy Build. 235 (2021), https://doi.org/10.1016/j.enbuild.2021.110730 110730.
[30] B. Yildiz, J.I.I. Bilbao, J. Dore, A.B.B. Sproul, Recent advances in the analysis of residential electricity consumption and applications of smart meter data, Appl. Energy. 208 (2017) 402–427, https://doi.org/10.1016/J.APENERGY.2017.10.014.
[31] J. Torriti, People or machines? Assessing the impacts of smart meters and load controllers in Italian office spaces, Energy Sustain. Dev. 20 (2014) 86–91, https://doi.org/10.1016/j.esd.2014.01.006.
[32] Demba Ndiaye, Kamiel Gabriel, Principal component analysis of the electricity consumption in residential dwellings, Energy Build. 43 (2-3) (2011) 446–453, https://doi.org/10.1016/j.enbuild.2010.10.008.
[33] F. McLoughlin, A. Duffy, M. Conlon, Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study, Energy Build. 48 (2012) 240–248, https://doi.org/10.1016/j.enbuild.2012.01.037.

[34] A. Kavousian, R. Rajagopal, M. Fischer, Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior, Energy. 55 (2013) 184–194, https://doi.org/10.1016/j.energy.2013.03.086.
[35] C. Beckel, L. Sadamori, S. Santini, Automatic socio-economic classification of households using electricity consumption data, Proc. Fourth ACM Work, Embed. Sens. Syst. Energy-Efficiency Build. (2013) 75, https://doi.org/10.1145/2487166.2487175.
[36] X. Tong, R. Li, F. Li, C. Kang, Cross-domain feature selection and coding for household energy behavior, Energy. 107 (2016) 9–16, https://doi.org/10.1016/j.energy.2016.03.135.
[37] G. Huebner, D. Shipworth, I. Hamilton, Z. Chalabi, T. Oreszczyn, Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes, Appl. Energy. 177 (2016) 692–702, https://doi.org/10.1016/j.apenergy.2016.04.075.
[38] J.L. Viegas, S.M. Vieira, R. Melício, V.M.F. Mendes, J.M.C. Sousa, Classification of new electricity customers based on surveys and smart metering data, Energy. 107 (2016) 804–817, https://doi.org/10.1016/j.energy.2016.04.065.
[39] J.P. Gouveia, J. Seixas, Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys, Energy Build. 116 (2016) 666–676, https://doi.org/10.1016/J.ENBUILD.2016.01.043.
[40] J.P. Gouveia, J. Seixas, A. Mestre, Daily electricity consumption profiles from smart meters - Proxies of behavior for space heating and cooling, Energy. 141 (2017) 108–122, https://doi.org/10.1016/J.ENERGY.2017.09.049.
[41] I. Laicane, D. Blumberga, A. Blumberga, M. Rosa, Comparative multiple regression analysis of household electricity use in Latvia: using smart meter data to examine the effect of different household characteristics, Energy Procedia. 72 (2015) 49–56, https://doi.org/10.1016/j.egypro.2015.06.008.
[42] G. Dane, L.G. Swan, A method for distinguishing appliance, lighting and plug load profiles from electricity 'smart meter' datasets, Energy Build. 134 (2017) 212–222, https://doi.org/10.1016/j.enbuild.2016.10.048.
[43] F. McLoughlin, A. Duffy, M. Conlon, A clustering approach to domestic electricity load profile characterisation using smart metering data, Appl. Energy. 141 (2015) 190–199, https://doi.org/10.1016/j.apenergy.2014.12.039.
[44] E. Hache, D. Leboullenger, V. Mignon, Beyond average energy consumption in the French residential housing market: A household classification approach, Energy Policy. 107 (2017) 82–95, https://doi.org/10.1016/j.enpol.2017.04.038.
[45] M. Azaza, F. Wallin, Smart meter data clustering using consumption indicators: Responsibility factor and consumption variability, Energy Proc. 142 (2017) 2236–2242, https://doi.org/10.1016/j.egypro.2017.12.624.
[46] Z.A. Khan, D. Jayaweera, M.S. Alvarez-Alvarado, A novel approach for load profiling in smart power grids using smart meter data, Electr. Power Syst. Res. 165 (2018) 191–198, https://doi.org/10.1016/j.epsr.2018.09.013.
[47] N.A. Funde, M.M. Dhabu, A. Paramasivam, P.S. Deshpande, Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data, Sustain. Cities Soc. 46 (2019), https://doi.org/10.1016/j.scs.2018.12.043 101415.
[48] C. Wang, Y. Du, H. Li, F. Wallin, G. Min, New methods for clustering district heating users based on consumption patterns, Appl. Energy. 251 (2019), https://doi.org/10.1016/j.apenergy.2019.113373 113373.
[49] S. Yilmaz, J. Chambers, M.K. Patel, Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management, Energy. 180 (2019) 665–677, https://doi.org/10.1016/j.energy.2019.05.124.
[50] K. Zhou, C. Yang, J. Shen, Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China, Util. Policy. 44 (2017) 73–84, https://doi.org/10.1016/J.JUP.2017.01.004.
[51] K. Li, Z. Ma, D. Robinson, W. Lin, Z. Li, A data-driven strategy to forecast next-day electricity usage and peak electricity demand of a building portfolio using cluster analysis, Cubist regression models and Particle Swarm Optimization, J. Clean. Prod. 273 (2020), https://doi.org/10.1016/j.jclepro.2020.123115 123115.
[52] S. Wang, H. Liu, H. Pu, H. Yang, Spatial disparity and hierarchical cluster analysis of final energy consumption in China, Energy. 197 (2020), https://doi.org/10.1016/j.energy.2020.117195 117195.
[53] S.P. Pieri, I. Tzouvadakis, M. Santamouris, Identifying energy consumption patterns in the Attica hotel sector using cluster analysis techniques with the aim of reducing hotels' CO2 footprint, Energy Build. 94 (2015) 252–262, https://doi.org/10.1016/j.enbuild.2015.02.017.
[54] P. Gianniou, X. Liu, A. Heller, P.S. Nielsen, C. Rode, Clustering-based analysis for residential district heating data, Energy Convers. Manag. 165 (2018) 840–850, https://doi.org/10.1016/j.enconman.2018.03.015.
[55] O. Laib, M.T. Khadir, L. Mihaylova, Toward efficient energy systems based on natural gas consumption prediction with LSTM Recurrent Neural Networks, Energy 177 (2019) 530–542, https://doi.org/10.1016/j.energy.2019.04.075.
[56] C. Laspidou, E. Papageorgiou, K. Kokkinos, S. Sahu, A. Gupta, L. Tassiulas, Exploring patterns in water consumption by clustering, Proc. Eng. 119 (2015) 1439–1446, https://doi.org/10.1016/j.proeng.2015.08.1004.
[57] G. Buttitta, W. Turner, D. Finn, Clustering of household occupancy profiles for archetype building models, Energy Procedia. 111 (2017) 161–170, https://doi.org/10.1016/j.egypro.2017.03.018.
[58] A.F. Meyabadi, M.H. Deihimi, A review of demand-side management: Reconsidering theoretical framework, Renew. Sustain. Energy Rev. 80 (2017) 367–379, https://doi.org/10.1016/j.rser.2017.05.207.
[59] A. Satre-Meloy, M. Diakonova, P. Grünewald, Cluster analysis and prediction of residential peak demand profiles using occupant activity data, Appl. Energy. 260 (2020), https://doi.org/10.1016/j.apenergy.2019.114246 114246.
[60] K. Zhou, S. Yang, Z. Shao, Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study, J. Clean. Prod. 141 (2017) 900–908, https://doi.org/10.1016/j.jclepro.2016.09.165.
[61] L. Wen, K. Zhou, S. Yang, A shape-based clustering method for pattern recognition of residential electricity consumption, J. Clean. Prod. 212 (2019) 475–488, https://doi.org/10.1016/j.jclepro.2018.12.067.
[62] KOM Ltd., (2018). https://kozpontiokosmeres.hu/ (accessed March 26, 2021).
[63] D.A.. Dillman, Mail and internet surveys: the tailored design method, 2000.
[64] Hungarian Central Statistical Office (KSH), Dwelling data in Hungary, (2020). https://www.ksh.hu/stadat_files/lak/hu/lak0001.html.
[65] Matlab R2017a, (2017). https://uk.mathworks.com/help/matlab/release-notes-R2017a.html.
[66] K. Li, R.J. Yang, D. Robinson, J. Ma, Z. Ma, An agglomerative hierarchical clustering-based strategy using Shared Nearest Neighbours and multiple dissimilarity measures to identify typical daily electricity usage profiles of university library buildings, Energy 174 (2019) 735–748, https://doi.org/10.1016/j.energy.2019.03.003.
[67] G. Chicco, R. Napoli, F. Piglione, Comparisons among clustering techniques for electricity customer classification, IEEE Trans. POWER Syst. 21 (2006) 933–940.
[68] P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, Introduction to Data Mining (Second Edition), Pearson, 2019.
[69] K. Le Zhou, C. Fu, S.L. Yang, Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation, Sci. China Inf. Sci. 57 (2014) 1–8, https://doi.org/10.1007/s11432-014-5146-0.
[70] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, Iñigo Perona, An extensive comparative study of cluster validity indices, Pattern Recognit. 46 (1) (2013) 243–256, https://doi.org/10.1016/j.patcog.2012.07.021.
[71] V. Vámos, L. Czétány, M. Horváth, T. Csoknyai, Gas Consumption Analysis for Educational buildings, Spec. Issue J. Heating, Vent. Sanit. 29 (2020) 327–331.
[72] J.N. Adams, Z. Deme Bélafi, M. Horváth, J.B. Kocsis, T. Csoknyai, How Smart Meter Data Analysis Can Support Understanding the Impact of Occupant Behavior on Building Energy Performance, A Comprehensive Review, Energies. 14 (2021) 1–23, https://doi.org/10.3390/en14092502.
[73] Hungarian Central Statistical Office (KSH), Settlement data Hungary, (2020). http://www.ksh.hu/apps/hntr.main?p_lang=HU.