Classification of daily electric load profiles of non-residential buildings

Mathieu Bourdeau, Philippe Basset, Solène Beauchêne, David da Silva, Thierry Guiot, David Werner, Elyes Nefzaoui

Classification of daily electric load profiles of non-residential buildings

Mathieu Bourdeau[1,2,*], Philippe Basset[2], Solène Beauchêne[3,4], David Da Silva[4,5], Thierry Guiot[4,5], David Werner[1], and Elyes Nefzaoui[2,4,**]

[1] *CAMEO SAS 55 Rue de Châteaudun, F-75009 Paris, France*
[2] *ESYCOM Lab, Univ Gustave Eiffel, CNRS, F-77454 Marne-la-Vallée, France*
[3] *EDF R&D, EDF Lab Les Renardières, F-77818 Moret sur Loing, France*
[4] *Efficacity, F-77447 Marne la Vallée Cedex 2, France*
[5] *Centre Scientifique et Technique du Bâtiment, Sophia Antipolis, France*
**Corresponding author: mathieu.bourdeau@esiee.fr*
***Contact author: elyes.nefzaoui@esiee.fr*

Abstract

We investigated clustering techniques on time series of daily electric load profiles of fourteen higher education buildings on the same campus. A k-means algorithm is implemented, and three different methods are compared: time-series features extraction with Manhattan distance and raw time series with Euclidian distance and Dynamic Time Warping. The impact of data characteristics with data collection time-steps and timeframes is studied using a database of more than 6,500 daily electric load profiles. We show that Euclidian distance applied to electric demand time series with three-month timeframes and ten-minute time-step provides the most consistent clustering results. In addition, useful insights are highlighted for non-residential buildings electric demand modeling and forecasting. Two groups of buildings can be distinguished regarding electric load profile patterns. On one hand, teaching, research, libraries, and gymnasium buildings show similar patterns distributed in two clusters corresponding to business days and closing days load profiles. On the other hand, campus office buildings present a larger number of clusters inconsistent with day-type dependent load profiles. A seasonal effect is also observed using six-month and one-year timeframes. Finally, a two-cluster distribution is obtained when aggregating all buildings load profiles.

Keywords

Clustering; daily load profiles; electric demand; non-residential buildings

Abbreviation

CVI	Cluster Validation Index	GF	Ground Floor
DLP	Daily Load Profile	TS-EUCL	Time series clustering with Euclidian distance
DTW	Dynamic Time Warping	TS-DTW	Time series clustering with Dynamic Time Warping
FB-MAN	Feature-based clustering with Manhattan distance		

1 Introduction

Reducing buildings energy consumption and related greenhouse gas emissions is one of the major challenges for research on the built environment. Indeed, buildings account for 29% of worldwide final energy consumption and 49% of the total electricity consumption [1], and this share is continuously increasing [2]. To tackle these issues, recent opportunities have risen in the development of smart infrastructures [3] and the significant role of smart meters deployment plans in the United States [4,5], in China [6] or in Europe [7], for instance. The growing availability of data collected from advanced metering infrastructure is therefore a strong asset for research and development towards a better and realistic understanding and modeling of buildings energy consumption.

To take the full benefit of the amount of collected data, the current trend in building energy modelling is shifting from traditional physics-based modelling [8] to data-driven methods [9]. To quantify the diversity of behaviors to be captured by such models, which may be translated into different sub-models, one can sort available datasets in different sub-sets exhibiting

similar characteristics. For this purpose, data clustering is widely used. Clustering methods have been implemented for a variety of purposes related to building energy consumption such as patterns recognition [10,11], abnormal energy behaviors identification [12], general building energy demand characterization [13,14], demand side management for industrial [15] and residential [16,17] sectors, building energy consumption [18] and peak demand [19] forecasting. These techniques are also used for various applications including the identification of priority targets for energy efficiency programs [20], the optimization of equipment sizing, energy storage, electric networks operation, renewables integration [21,22] and commercial offers [23,24]. Studies have mainly covered residential households and then mixed industrial and commercial buildings as highlighted in [19]. Other non-residential buildings such as education, research or office buildings have more seldomly been considered [11,25].

Hence, clustering applications for building electric demand analyses have been increasingly addressed. Nevertheless, a majority of studies use very large databases comprising electricity demand data of several hundred [12,16] or even thousands of buildings. They often refer to public database with the PecanStreet Database [26] using around 600 buildings [27] and the Irish Commission of Energy Regulation [28] with about 4,000 buildings [13,17,29]. Indeed, as the amount of available data is constantly increasing, it provides interesting insights for large-scale electricity demand analysis and related applications, and enables the implementation of robust algorithms [23]. However, such studies focus more on the performance of clustering methods and the extraction of large-scale trends. The amount of processed data prevents more detailed analyses at building- or district-scale to provide a deeper physical understanding of buildings electric demand behaviors. Also, data collection timeframes are often limited to short periods of time with a year of data [12,14] or less [13,27], and then divided in few-month-long sub-datasets. Although they might lead to

accurate classifications, short data collection timeframes mask relevant information regarding building electricity demand such as seasonal effects. Finally, the comparison of clustering methods has been extensively studied. For instance, performances of k-means and hierarchical clustering algorithms have been investigated by Quintana et al. [12], Satre-Meloy et al. [19], Chicco et al. [23] who also compared with fuzzy k-means and "follow the leader" algorithms and Xu et al. [27] who tested adaptive k-means and symbolic aggregate approximation (SAX) methods as well. However, the impact of collected data properties on the clustering methods performance and clustering results, apart from feature engineering (i.e. data time-step and timeframe collection), has been less considered.

Therefore, we address these challenges by proposing a study of daily load profile (DLP) classification for electric demand pattern identification for fourteen higher education buildings located on the same campus, in Paris eastern suburb. Considered methods are tested with time series of buildings electric demand collected between December 2014 and April 2019 using advanced metering tools [30] and resulting in a database of more than 6,500 daily electric load profiles. A k-means algorithm is used with three different approaches. Two types of inputs are investigated. Feature-based clustering is performed using Manhattan distance metric (FB-MAN). Raw time series clustering is computed comparing two distance metrics with Euclidian distance (TS-EUCL) and Dynamic Time Warping (TS-DTW). A comparative analysis is first performed on two buildings considered in previous studies [31,32] using the three clustering methods with different input data properties (observation time-steps and timeframes). It highlights the respective accuracy of the algorithms depending on input data characteristics. It also provides physical understanding regarding the buildings electric demand with a two to three-cluster day-type-based classification pattern and an occupancy-related seasonal effect. Clustering tests are then generalized to the whole building stock, leading to identify two distinctive building electric demand classification patterns with

campus office buildings in one group and teaching/research buildings, libraries, and gymnasium in another group. Finally, the general day-type-based classification is confirmed when considering aggregated DLP of all buildings which exhibit two-cluster distribution separating business days from university closing days. The paper is organized as follows: in Section 2 the different buildings considered in this case study are presented. Section 3 introduces the methods for data collection, pre-processing, and clustering. Obtained results are reported and discussed in Section 4.

2 Case study

In the present work, fourteen non-residential buildings and groups of buildings are considered and further referred to as B1–B12. All buildings are individual units except for three teaching and research buildings grouped together (B6), and for which load data are collected from one single electricity meter. Two of the buildings (B11 and B12) were already investigated in detail in previous studies [31,32]. Buildings are located on a same university campus in Paris eastern suburb, France. The fourteen buildings cover different common activities of an academic campus including administrative offices, classrooms, amphitheaters, research, but also a library, a gymnasium and rooms for student organizations. General features of the buildings are presented in Table 1.

Building number	**Building type**	**Specific activity**	**Net floor area (m^2)**	**Floors**	**Annual electric energy consumption – calendar year 2019 (kWh)**	**Annual surface electric energy consumption density (kWh/m^2)**	**Contracted power (kW)**	**Surface contracted power density (W/m^2)**
B1	Office building	Examination center	1,233	GF+6	94,989	77.0	60	48.7
B2	Library	/	8,799	GF+2	463,969	52.7	430	48.9
B3	Teaching and Research	Economics, humanities and social sciences	11,443	GF+2	389,069	34.0	138	12.1
B4	Office building	Business incubator	/	GF+1	173,276	/	136	/
B5	Teaching	Science and	25,100	GF+3	2,139,075	85.2	1420	56.6

	and Research	Technology						
B6	Teaching and Research	Science and Technology	16,800	GF+3	680,316	40.5	260	15.5
B7	Gymnasium	/	2,002	GF	42,085	21.0	36-42	17.9-21.0
B8	Teaching and Research	Science and Technology	10,428	GF+2	274,533	26.3	120	11.5
B9	Office building	Hosting students' activities	1,945	GF+1	66,445	34.2	120	61.7
B10	Library	/	6,360	GF+3	851,432	133.9	160-230	25.2-36.2
B11	Teaching and Research	Art and Humanities	10,343	GF+5	645,129	62.4	250	24.2
B12	Teaching and Research	Science and Technology	30,580	GF+4	4,622,912	151.2	950	31.1

Table 1 – General features of the fourteen building case studies

3 Methods

3.1 Data Collection and pre-processing

3.1.1 Data collection

Available data are electric load time series data collected with meters and provided by a distribution system operator. It includes the instantaneous apparent power, active power, reactive power and voltage monitored with a ten-minute time-step. In the present work, only the active power is considered. The collection timeframe varies from a building to another. Electricity demand collection started on 12/07/2014 at the earliest, for B12, and at the latest on 05/01/2018 for B1, B2, B3, B4, B7, B8, B9 and B10. Details on the data collection timeframes are given in Table 2.

3.1.2 Software

Data are formatted and cleaned using Python 3.7.3 [33]. Data features are extracted using Numpy 1.16.4 package [34]. Clustering is performed using R 3.6.1 software [35] and NbClust 3.0 package [36]. The NbClust package uses the calculation of 30 cluster validation indices

(CVI) to determine the optimal number of clusters and distribution of data for a given dataset. TS-DTW clustering is also performed using Nbclust along with TSClust 1.2.4 package [37], a package for dissimilarity measurement between time series to perform time series clustering.

3.1.3 Data pre-processing

Raw power demand data are provided on a single column format with each line corresponding to one timestamp for the selected timeframe. Data formatting is performed by extracting daily load profiles (DLP). DLPs are then the input dataset for clustering algorithms, with the number of lines corresponding to the number of days in the selected timeframe and the number of columns corresponding to the number of data points in each day. Raw data are downloaded at ten-minute, thirty-minute and hourly time-step. Thirty-minute and hourly data are the average of the previous three and six ten-minute data points, respectively.

Data cleaning is performed to identify low-quality data including missing data and data collection failures. The latter can include outliers such as negative or overly-high power demand as well as error messages. Single values are considered overly-high when they exceed ten times the average daily power demand – they usually result from power failures leading to false meter readings. Outliers are treated using mean substitution method [38]: they are removed and replaced by the average of the nearby data points if they do not exceed an hour of data. Otherwise the corresponding day is not used for classification. Missing data are also processed depending on the number of missing data points. When the gap exceeds a day of data, it is related to electric interventions in the buildings, for instance for construction work. Then the corresponding timeframe is not used for classification. Otherwise, few-hour-long missing periods are due to temporary power failures. They are not removed nor corrected. They are the representation of real unexpected events happening in the studied buildings. Hence, it is relevant to keep these profiles to test clustering algorithms for anomalous days

151 detection. Missing data points are replaced with zero values for the algorithm to account for
152 these data (otherwise these specific DLPs would not be considered by the algorithm).

Building number	**Beginning date of data collection**	**End date of data collection for the present study**	**Amount of available data (at 10-minute time-step)**	**Amount of low-quality data**	**Number of available DLP after data cleaning**
B1	05/01/2018 00:00	04/30/2019 23:50	52,560 data points	0 data points – 0%	365 DLP
B2	05/01/2018 00:00		52,538 data points	22 data points – 0.04%	365 DLP
B3	05/01/2018 00:00		52,544 data points	16 data points – 0.03%	365 DLP
B4	05/01/2018 00:00		52,560 data points	0 data points – 0%	365 DLP
B5	06/01/2017 00:00		100,646 data points – 699 days	10 data points – 0.01%	699 DLP
B6	06/01/2017 00:00		96,147 data points – 699 days	4,509 data points – 4.48%	668 DLP
B7	05/01/2018 00:00		52,560 data points	0 data points – 0%	365 DLP
B8	05/01/2018 00:00		52,516 data points	44 data points – 0.08%	365 DLP
B9	05/01/2018 00:00		52,560 data points	0 data points – 0%	365 DLP
B10	05/01/2018 00:00		52,547 data points	13 data points – 0.02%	365 DLP
B11	01/01/2017 00:00		122,358 data points	42 data points – 0.03%	850 DLP
B12	12/07/2014 00:00		207,421 data points	23,843 data points – 10.3%	1,440 DLP
Total	/	/	946,957 data points	28,499 data points – 3.00%	6,577 DLP

153 Table 2 – Details of the collected data

154 3.2 Clustering methods

155 3.2.1 General process

156 The first step of the clustering process includes cleaning of raw time series and preparing
157 datasets with specific timeframes and time-steps. Selected features are also extracted. Then
158 clustering is performed with NbClust package [36]. This package computes thirty CVIs that
159 are cluster evaluation metrics. Each index is computed to assess the optimal number of

clusters and the corresponding optimal data distribution for each cluster. A majority rule, implemented in NbClust algorithm, selects the results: the final data distribution and optimal number of clusters are designated by the highest number of CVI indicating the same results. K-means algorithm [39] is used for clustering, and results are computed using three methods. The first method is based on five features as input data extracted from raw time series and using the Manhattan distance metric (FB-MAN). It is compared with two other approaches using time series as input data and with different time-steps: ten minutes, thirty minutes and hourly. The second method is based on Euclidean distance metric (TS-EUCL) while the third method is based on DTW (TS-DTW).

The tested number of clusters is bound between two and six to restrict the computational time and to avoid the distinction between similar DLPs exhibiting very small differences. Based on prior studies on similar buildings [31,32] testing up to a six-cluster distribution can be considered reasonable to synthetize the main significant DLPs of the different buildings and eventually to identify unusual electric demand patterns. Also, because of the random initialization of k-means in NbClust and to ensure the validity of clustering results, a loop is implemented to run the algorithm one hundred times with random initialization. The selected results, i.e. the optimal number of cluster and DLP distribution, are those provided by the largest number of runs and at least by 80% of the total number of runs.

Also, the NbClust process is modified to add a complementary step after the calculations of CVIs. This step checks for results designating two different optimal number of clusters or when a difference of one CVI is given between the optimal number of clusters and the second ranking solution. Indeed, both cases could present relevant data distributions that would require further investigation before a final selection. Moreover, after the choice of the optimal number of clusters, distributions provided by different CVIs are compared to ensure they are identical. Afterwards, the building activity schedule based on worked days, weekends and

vacations (when the information is available) is provided along with the DLP distribution. DLPs are plotted for each different cluster. DLPs in each cluster are displayed in grey and the average DLP profile of each cluster is added with a bold black curve to highlight the main shape of DLPs grouped together. The average profile is obtained by averaging the power demand of all DLPs for each time-step. The cluster distribution is also mapped on a calendar for large timeframes over six months to improve the readability of the DLP classification.

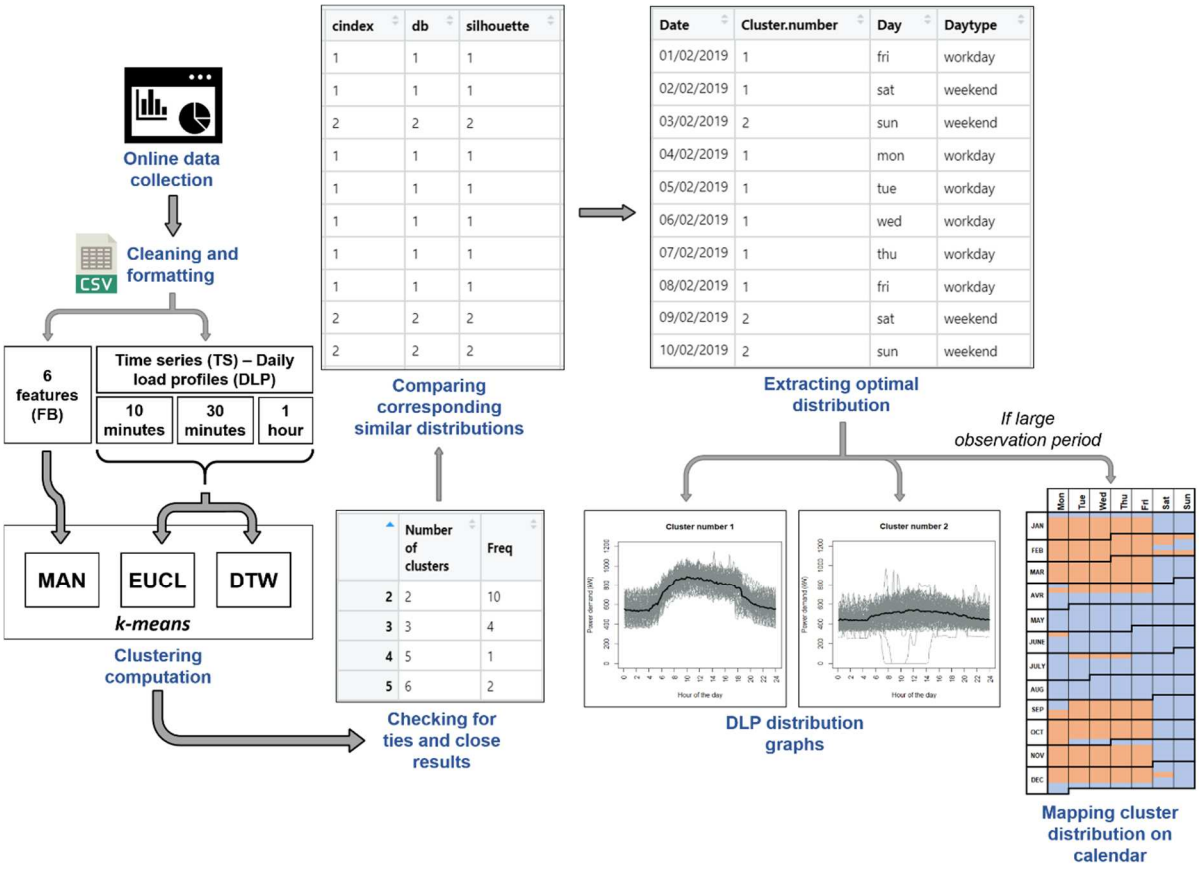


Figure 1 – General clustering process

3.2.2 Clustering algorithm

Clustering can be performed with various methods. Three of the most used classification algorithms for building energy consumption applications [39] are self-organizing maps (SOM) [18], hierarchical clustering [19] and k-means [15,16]. For the latter, several modified algorithms are found, such as adaptive k-means [27], fuzzy k-means [23] and k-means++ [11] as well as reminiscent techniques including k-medoids [13,40] and k-shape [12,25]. Finally,

198 other methods are considered as well including symbolic aggregate approximation (SAX)
199 [17,27], finite mixture models (FMM) [29] and “Follow the leader” clustering algorithm
200 [10,23].

201 In the present study, a k-means algorithm is used. K-means is a simple but efficient and
202 versatile clustering technique [16] that is the most used algorithm for performance analysis of
203 non-residential buildings [41]. It is an iterative unsupervised non-hierarchical classification
204 method which divides a set of data into k different clusters, with k being user-defined. To
205 create the clusters, k initial data points are first randomly selected as centroids that is the
206 center of a cluster. Then the similarity between each new data points and centroids is assessed
207 and data points are assigned to the cluster whose centroid is the nearest. The notion of
208 similarity is assessed using a distance calculation. In the present case, three distances are
209 used. Feature-based clustering is computed with Manhattan distance. Raw electric power
210 time-series clustering is performed with Euclidian distance and Dynamic Time Warping
211 method. Then, after each iteration clusters centers are recalculated considering the added data
212 points and the distance between each data points and the newly calculated centroids is
213 assessed again. The clustering process is iterated until no data points are reassigned to new
214 clusters. Hence, k-means algorithm is selected among other algorithms particularly as it is
215 particularly adapted to the present case study. Indeed, the goal is to group DLPs around a
216 typical mean DLP and then identify the prominent DLP shapes for a given building (i.e. what
217 is presented in DLP distribution graphs), which is exactly what is provided by k-means.

218 3.2.3 Feature-based clustering

219 Features extraction is a characterization of raw time-series data which aims to reduce the size
220 of datasets while capturing key features of each time-series. If appropriately engineered,
221 extracted features enable computational time decrease, and eventually provide a physical

meaning to identified clusters. The main drawback of features extraction is the loss of relevant information with the reduction of datasets size. These features depend on the considered end applications that require data classification [42]. For feature-based clustering of power demand time-series, both physical and statistical features have been reported [43]. In the present study, both types of features are used. The following physics-based features are considered: maximum and minimum daily power demand, daily magnitude (the difference between the two latter features) (Eq.(1)) and daily electric energy consumption (Eq. (2)). Statistical features include standard deviation (Eq. (3)). Hence, five features are extracted for each load curve.

$$Magnitude = Max(y_1, \dots, y_{t=N}) - Min(y_1, \dots, y_{t=N}) \quad (1)$$

$$Energy = \sum_{t=1}^{N} y_t \,/(N/24) \quad (2)$$

$$Mean\ deviation = \sum_{t=1}^{N} |y_t - \bar{y}| \,/N \quad (3)$$

where y_t the value of the time-series at time *t*, and *N* the number of data points in a day (N=24, 48 or 144 with hourly, thirty-minute and ten-minute time-steps respectively).

Because the five selected features present different magnitudes, a normalization procedure is used prior to clustering. It rescales all the features and prevents the k-means algorithm from being driven by the variable showing the highest values (here the daily electric energy consumption). For this purpose using min-max normalization (Eq. (4)).

$$y_t' = \frac{y_t - \min(y_t)}{\max(y_t) - \min(y_t)} \quad (4)$$

where y_t is the value of the time-series at time *t* and y_t' the normalized value of y_t using min-max normalization.

As the different features have different physical meanings, feature-based clustering is computed using Manhattan distance metric (FB-MAN).

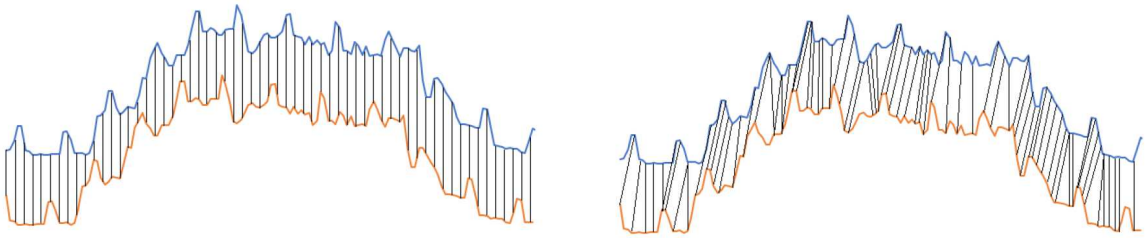
$$MAN = \sum_{t=1}^{N} |x_t - y_t| \quad (5)$$

3.2.4 Raw time series clustering

Unlike FB-MAN, raw time series clustering directly relies on pre-processed electric load time-series data [42] without prior extraction of features, that is without any loss of information with respect to pre-processed data. These methods compare daily load curve shapes using distance calculations. Various distances have been reported in the literature such as Euclidean [16,19], Manhattan [44], Chebyshev [19] or DTW [40]. In the present paper, two different distance metrics are used. The first is the Euclidean distance (TS-EUCL), a classical distance calculation to measure the similarity between two vectors or sets of data [42]. Considering two time series x_t and y_t, the Euclidean distance can be calculated as follows:

$$EUCL = \sqrt{\sum_{t=1}^{N} (x_t - y_t)^2} \quad (6)$$

The second method used is Dynamic Time Warping (TS-DTW) [45]. While Euclidean distance linearly matches the data of the two compared times series for every data point at the same time-step (Figure 2a), DTW rather stretches and compresses the time axis to align the time series so that an optimal alignment is found and the distance measure is minimized (Figure 2b). In the present case, with a fine time-step inducing a high variability of DLPs, DTW can be very useful to capture DLPs with similar shapes but exhibiting small phase shifts due to non-synchronized and shifting electric energy consuming events. The detailed theory of DTW computation falls out of the scope of the present study, but interested readers can refer to [45].



(a) (b)

Figure 2 – Representation of Euclidian (a) and Dynamic Time Warping (b) matching

3.2.5 Clustering validity indices

Clustering validity indices (CVIs) are used to determine the optimal number of clusters and clusters assignment in a classification problem. In the initial clustering process in NbClust calculates thirty different CVIs to assess the optimal number of cluster and data partition. However, several rounds of tests show that for small timeframes of one month, calculations of several indices do not converge. Consequently, in the present study, we considered only seventeen indices which systematically converge for any type of input data: Krzanowski- Lai index (kl), Calinski-Harabaszch index (ch), Hartigan index, C-Index (cindex), Davies-Bouldin index (db), Silhouette index, Ratkowsky-Lance index (ratkowsky), Ball-Hall index (ball), Point-biserial index (ptbiserial), gap statistic index (gap), McClain-Rao index (mcclain), Gamma index, Gplus index, Tau index, Dunn index, SD index and SDbw index. Among these, Calinski-Harabaszch [25], Silhouette [16,25], Davies-Bouldin [17,25], gap statistic [46] and Dunn [25] indices are commonly used CVIs, either alone or compared with each other to evaluate clusters.

4 Results and discussions

Two key information are retrieved with buildings DLP classification: the number of clusters and, since we consider physical data, the physical explanation of the DLPs distribution over

277 the different clusters. The number of clusters provides the number of typical building electric
278 demand patterns. Therefore, it indicates the number of sub-models that should be developed
279 to enhance modelling accuracy and to capture the diversity of electricity demand profiles. The
280 DLP distributions over the different clusters provide insights on the physical explanatory
281 variables of the different electricity demand patterns that can be used in predictive models.

282 In the present study, a large amount of data is available at different time-steps. Timeframes
283 and time-steps are two key data characteristics expected to have an impact on clustering
284 results, both for the number of clusters and the DLP distributions. A comparison is then
285 performed between different combinations of timeframes and time-steps with the three
286 proposed approaches, FB-MAN, TS-EUCL and TS-DTW using data from B12. Indeed, from
287 previous studies [31,32], more meta-data are available for this building compared to other
288 case studies, with respect to occupancy, activity schedules, equipment, operation and other
289 characteristics. They are used to analyze and compare cluster distribution results. Similar tests
290 are conducted on B11 as well to validate the obtained results. For the sake of clarity, the
291 following subsections present the clustering results for B12 only.

292 4.1 First insights from a sample building

293 The comparison is performed between ten-minute, thirty-minute, and hourly time-steps for
294 timeframes of one month, three months, six months and one year of data between 05/01/2018
295 and 04/30/2019. Timeframes contain school vacation days (for students only, in that case
296 building are open with partial staff occupancy), weekends, national holidays, annual closing
297 days and normal business days. Comparative results are summarized in Table 3 and are
298 described in detail with the type of days in each cluster in Table 7 in Appendix 1.

Timeframe	**Time-step**	**B12**		
		FB-MAN	**TS-EUCL**	**TS-DTW**

1 month *01/01/2018 to 01/31/2018*	10-minute	3	3	4
				3
	30-minute	3	3	2
	Hourly	3	3	2
3 months *01/01/2018 to 03/31/2018*	10-minute	3	3	2*
	30-minute	2*	3	2*
	Hourly	2**	3	2**
6 months *01/01/2018 to 06/30/2018*	10-minute	3	2	2**
	30-minute	2*	2	2*
	Hourly	2**	2	2*
1 year *01/01/2018 to 12/31/2018*	10-minute	3	2*	2*
	30-minute	2*	3	2**
	Hourly	2**	2**	3

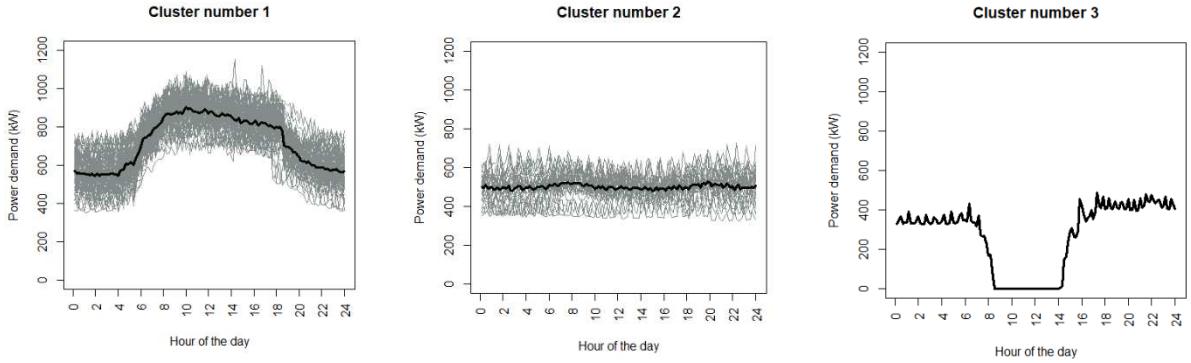
Table 3 – Comparative clustering results for B12 – Number of clusters marked with stars in superscript differentiate the results with different DLP distributions for a same number of clusters: one star highlights a given DLP distribution and two stars highlight a different DLP distribution

4.1.1 Timeframes, day-type-based classification and seasonal effect

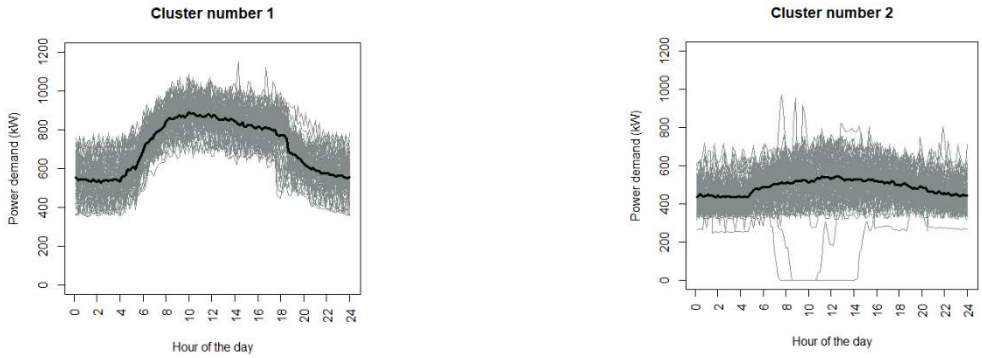
A main data characteristic is the timeframe, i.e. the duration of observation of electric load, used for clustering which can range, in this case, from one day to several years. With the amount of available data for the present study, a large diversity of DLPs can be expected. This diversity may lead to different building electric demand signatures due to various building use scenarios with respect to planning, occupancy or weather conditions. The number of clusters strongly depends on this diversity, hence on the timeframe size.

The effect of the timeframe can be clearly observed in Table 3. Indeed, the shorter the timeframe, the higher the number of clusters and the more detailed the results. With a one-month timeframe, TS-EUCL and FB-MAN both highlight three-cluster distributions: cluster 1 groups business days, cluster 2 groups closing days (weekends and Christmas vacations) and cluster 3 highlights an outlying profile. The same distribution is also found with TS-EUCL and three-month timeframe (Figure 3a), and with FB-MAN using a ten-minute time-step with

one-month timeframe. In addition to the above-mentioned general trend, some exceptions should be noted depending on the method. TS-DTW does not provide the same classification than the other two methods for one-month timeframe. It identifies the outlier only in the four-cluster distribution and separates Christmas vacations in two clusters for a ten- minute time-step due to small variations in the DLP shapes, or it groups Christmas vacations with weekends for larger time-steps. Also, TS-DTW and FB-MAN (except with ten-minute timeframe) do not highlight the outlying profile for three-month timeframe.



(a)



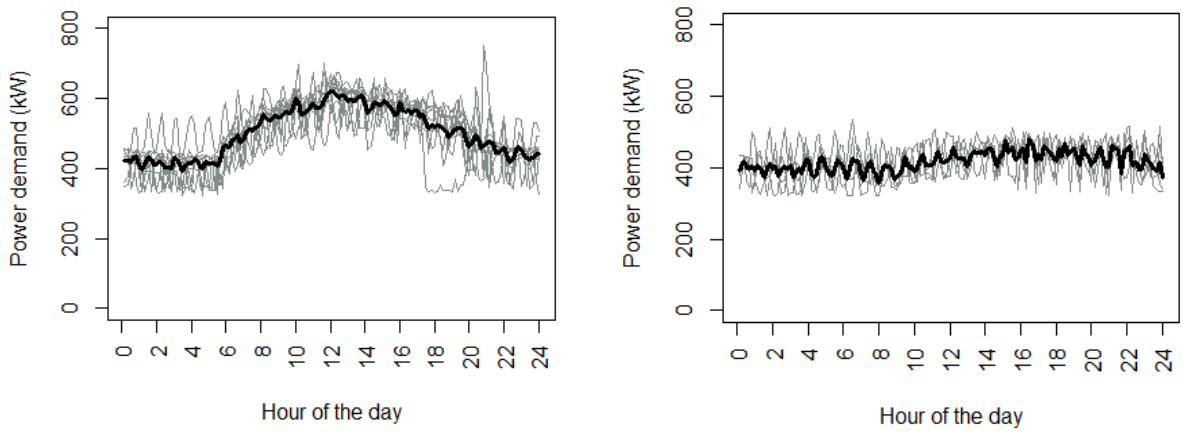
(b)

Figure 3 – Comparison of clustering results for B12 using TS-EUCL with 10-minute time-step data: three clusters (cluster 1: business days, cluster 2: closing days and cluster 3: an outlier) with a three-month timeframe (a) and two clusters (cluster 1: business days and cluster 2: mixed business and closing days) with a six-month timeframe (b) – the bold black curve highlights the average DLP for each cluster

Larger timeframes of six-month and one-year lead to a two-cluster or three-cluster DLP distribution. For the former, the first cluster groups business days. The second cluster groups weekends and other closing days such as Christmas vacations, national holidays, students vacation as well as the whole period from April to September, the beginning of the new academic year (or to the end of June for six-month timeframes) (Figure 3b). FB-MAN clustering exhibits slightly different results, with a three-cluster distribution for ten-minute time-step and two-cluster distribution for thirty-minute and hourly time-steps. The latter two time-steps show a similar DLP distribution to TS-EUCL and TS-DTW. However, ten-minute time-step yields different clustering results with three clusters obtained (Figure 6). The first cluster groups business days from January to the first two weeks of April and from September to December. The second cluster, as for TS-EUCL and TS-DTW groups weekends, Christmas vacations, some closing days (vacations and national holidays) and the whole summer period from the middle of April to the end of August. Then, the third cluster groups winter vacations in February and a few closing days when the activity and electric power demand are lower but not as high as in cluster two. Nevertheless, the seasonal effect is still observed as the difference between business days and closing days between April and September (Figure 4) is not highlighted in the clustering results.

With six-month and one-year timeframes, the day-type-based classification becomes less obvious. Indeed, for all clustering approaches, the two types of days are merged in the second cluster. From mid-July to the beginning of the new academic year, the buildings are closed for the university summer break, and from April to mid-July buildings are opened with lower activity due to the limited number of students and staffs. There is then a difference between business days and closing days (Figure 4). Nevertheless, both types of days are grouped with winter closing days – during which HVAC auxiliary systems are constantly activated to maintain the temperature of the buildings, causing a higher electricity demand than for

summer weekends. A seasonal effect is observed and can be explained by the lower number of occupants in the buildings before the summer closing days and because building occupancy is the main electricity demand driver [32]. This seasonal effect is of peculiar importance since it has a significant impact on daily electric demand with a lower electricity demand in summer compared to winter business days. It then results in a cluster of summer business days together with autumn and winter weekends. Therefore, it should be taken into account for modeling and forecasting applications.



(a) (b)

Figure 4 – DLPs from B12 for business days (a) and closing days (b) from 06/11/2018 to 06/24/2018

4.1.2 Time-steps

Data acquisition time-step is a key parameter for electric power demand data. Different time-steps offer different types of information on buildings electricity demand, relevant for different purposes. A fine sub-hourly temporal granularity can be used to detail the different electric energy uses and spot specific electricity demand patterns, while it also induces more data to store and process. A larger time-step, hourly or above, loosens constraints regarding data storage processing and provides aggregated electricity consumption information. However, it induces a significant information loss regarding fine demand features useful for

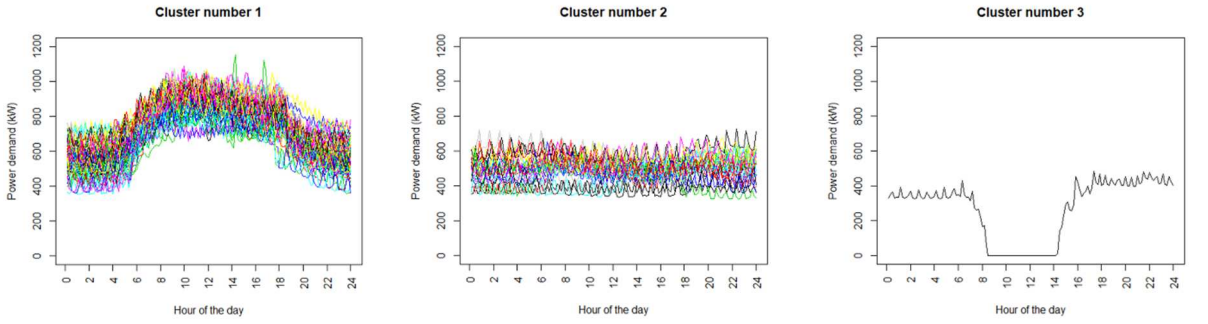
applications which strongly depend on power demand dynamics and such as renewable energy self-consumption or demand-response.

Overall, the different clustering tests show that the time-step affects the clustering results differently with respect to the clustering method. For TS-DTW, the impact of the time-step is significant. A sub-hourly time-step results in more detailed DLP classifications (Table 7). More specifically, with a one-month timeframe, TS-DTW is detecting small variations in the DLP of Christmas vacations and it separates the profiles in different clusters. Also, with a ten-minute time-step for larger timeframes, clustering results are the same as with other methods. However, the main issue with TS-DTW is the inconsistence of the results: a change of time-step induces at least one difference in data distributions or in the number of clusters for each timeframe.

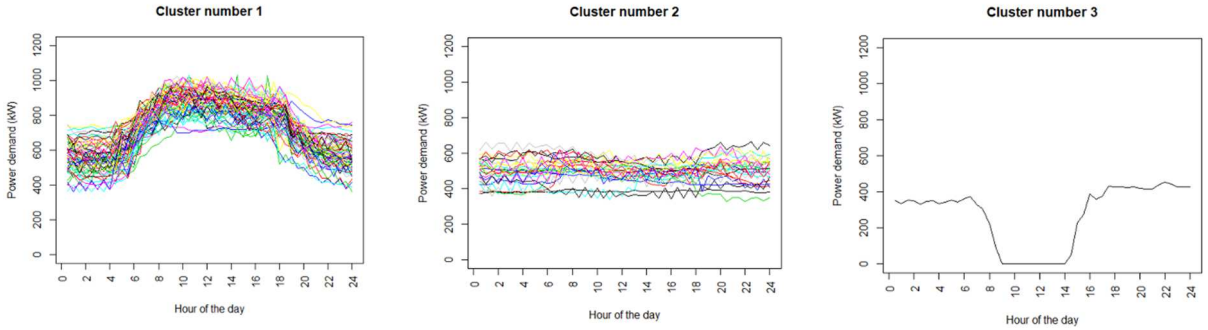
FB-MAN clustering is also affected by a change of time-step. Indeed, it leads to a change regarding the number of clusters and DLP distributions for timeframes larger than a month. Thirty-minute and hourly time-steps systematically lead to two clusters but with different DLP distributions (c.f. Table A.1). However, ten-minute time-step shows a three-cluster classification. For one-month and three-month timeframes, it successfully isolates the outlying profile. For larger timeframes, FB-MAN is affected by the seasonal effect previously described as it gathers business days from April the end of August together with weekends of the whole year and other closing days. Nevertheless, it also displays in the third cluster some specific days with lower activity and electric power demand profiles such as winter vacation days in February.

Then, TS-EUCL shows no or very little differences in the clustering results, neither in the number of clusters or the distributions. For one-month and three-month timeframes, the classification is the same for all three time-steps with respect to the types of days (Figure 5). For six-month and one-year timeframes, results are similar as well with the same seasonal

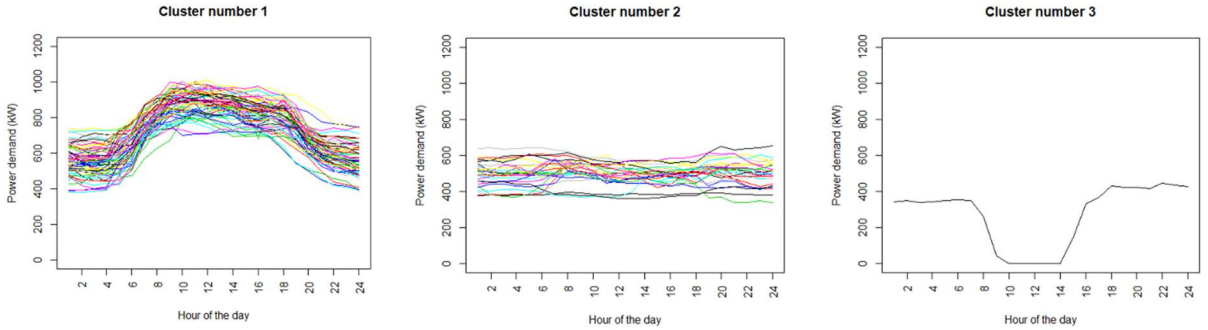
effect. This confirms that the daily electric energy consumption of business days from this time of the year is similar to the daily electricity consumption of winter closing days. One exception must be noted for one-year timeframe with thirty-minute data: instead of a two-cluster distribution, a three-cluster distribution is proposed.



(a)



(b)



(c)

Figure 5 – Comparison of DLP distributions in three clusters (cluster 1: business days, cluster 2: closing day and cluster 3: an outlier) for B12 using TS-EUCL with three-month timeframe and 10-minute time-step (a), 30-minute time-step (b) and hourly time-step (c)

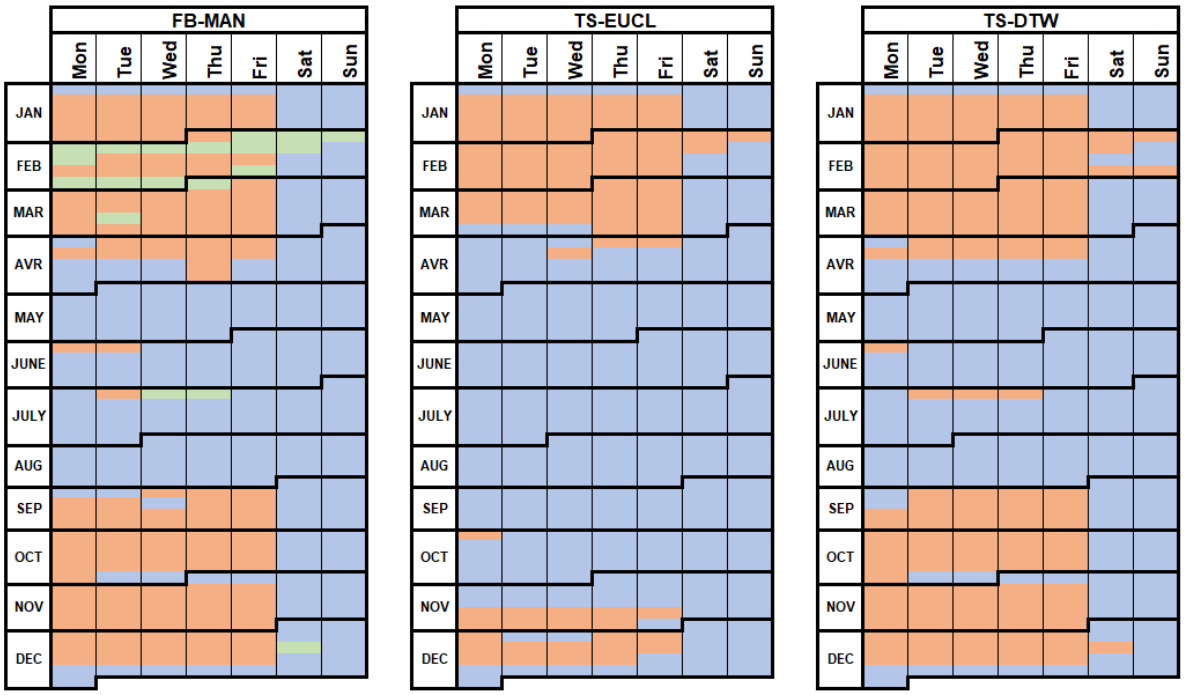
4.1.3 Performance comparison of the three methods

As highlighted in the previous subsections, the different clustering methods respond differently to a change of time-step or timeframe. Twelve different input data configurations are tested for each clustering methods by varying the timeframes and the time-steps. FB-MAN is quite sensitive to a change of time-step (Table 3). Specifically, ten-minute time-step provides additional information in the classifications. These are based on the types of days (business days or closing days) and the outlying profile (presented in cluster 3 of Figure 3a) is highlighted with a one-month for all three time-steps and three-month timeframes only with ten-minute time-step. The seasonal effect is present in large timeframes as for raw time series clustering. However, again with ten-minute time-steps, FB-MAN highlights an intermediate cluster grouping a few vacation days in the winter season, whose electric power demand is lower than usual business days but higher than closing days. Nevertheless, the sensitivity of FB-MAN clustering method to the time-step of input data may question the reliability of the results.

TS-DTW also provides DLP distributions with respect to the building activity overall. TS-DTW even shows more accurate clustering results for one-year timeframe than TS-EUCL and equivalent to FB-MAN (Figure 6). However, it is also very sensitive to the time-step. With different time-steps, results give different number of clusters but also different distributions for a same number of clusters. Hence, it is difficult to assess the optimal input data configuration and to rely on TS-DTW.

Then for TS-EUCL, one-year timeframe results are inconsistent. Using sub-hourly time-steps with one-year timeframe unexpectedly adds the whole period from September to the middle of November to the cluster with the seasonal effect previously described. As it is the beginning of the new academic year, this period has the highest occupancy. Hence, the DLP classification should cluster business day load profiles with other business days of the year,

427 which is not the case with these tests. However, TS-EUCL is the only method that highlights
428 the outlying profile using a three-month timeframe and all three time-steps. Furthermore,
429 unlike FB-MAN, using raw time series for clustering does not result in a loss of information
430 on the building electricity demand. In terms of modelling and forecasting applications, this is
431 a significant asset to consider large-enough observation periods of the building electricity
432 demand. For this reason, TS-EUCL is the method selected for the generalization to the whole
433 building stock.



434 Figure 6 – Comparison of cluster distributions over one-year timeframes for B12 with FB-
435 MAN, TS-EUCL and TS-DTW using a 10-minute time-step – Each cell corresponds to one
436 day and each line represents one week – Cluster 1 is figured in light brown, cluster 2 is
437 showed in light blue and cluster 3 is displayed in light green

4.2 Generalization to the whole building stock

4.2.1 Clustering on individual buildings: specificities of office buildings

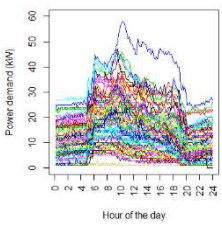
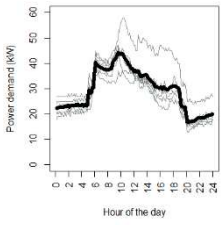
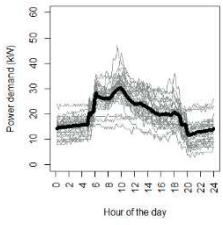
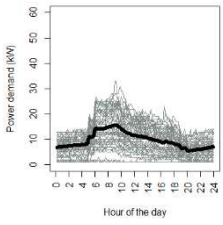
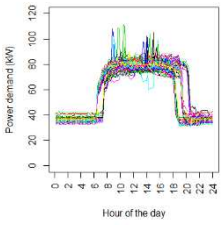
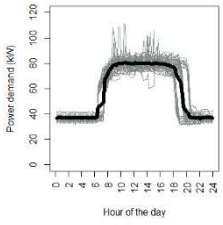
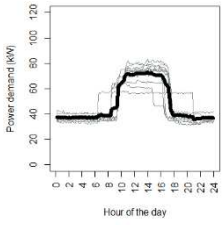
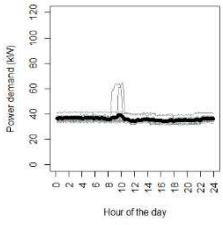
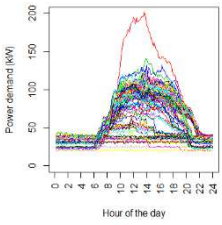
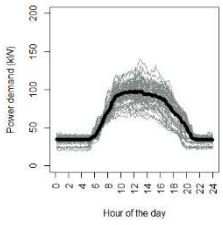
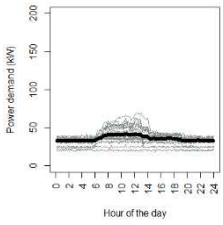
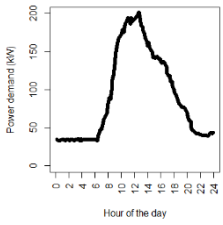
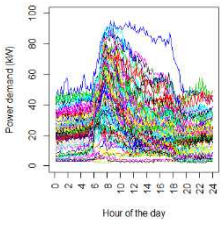
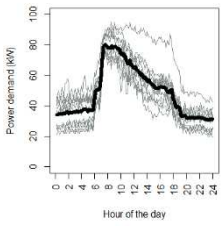
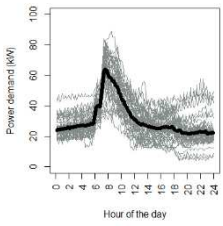
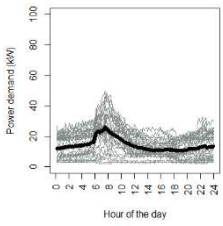
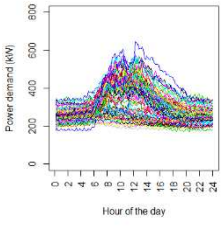
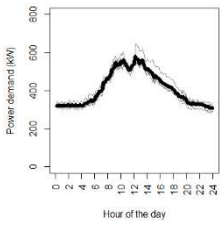
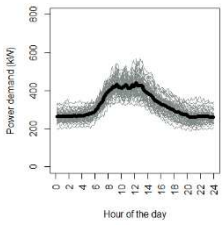
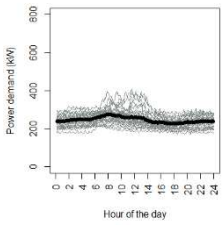
After first trials with B12 data and discussion of obtained results, TS-EUCL is applied on all building to investigate the main electricity demand patterns that can be extracted from DLP classifications. At this step, datasets of each building are considered separately. A three-month timeframe is used for all buildings with a data sample from 02/01/2019 to 04/31/2019. Indeed, an observation period of three months provides a reasonable amount of data with diverse building usages. Also, a ten-minute time-step is used as it does not aggregate information regarding the building electricity demand. On the other hand, after first trials on B12, TS-EUCL has proven to be the only method to provide a consistent day-type-based DLP classification with this timeframe and it was not affected by a change of time-step. Clustering results for all buildings are presented in Table 6. We show in column 2 all DLPs for the considered timeframe for each building and in the following columns the DLPs distributions are given along with the number of business days and closing days for each cluster.

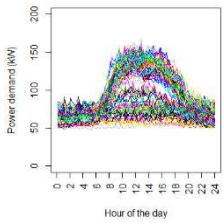
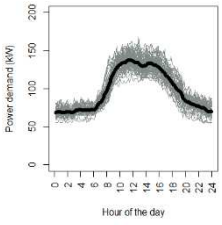
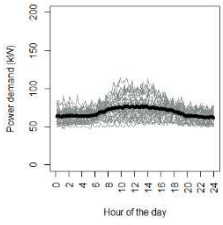
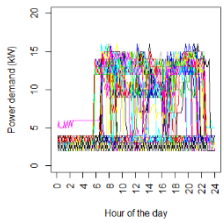
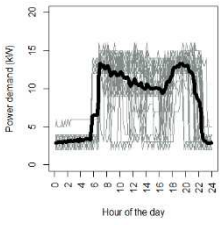
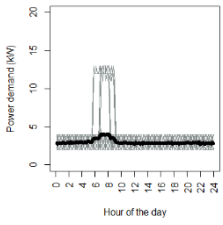
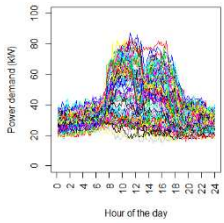
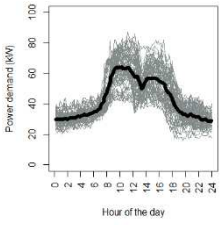
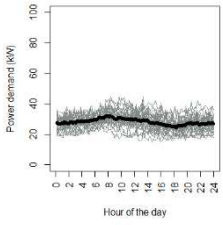
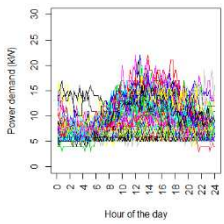
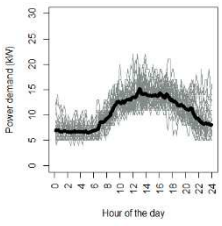
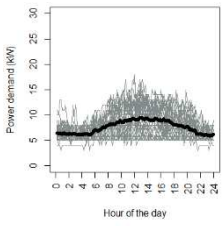
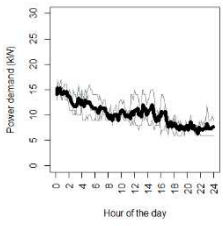
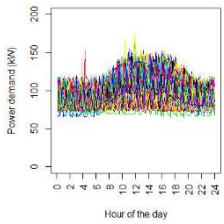
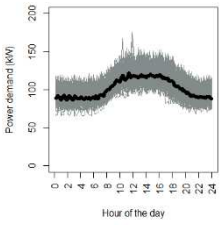
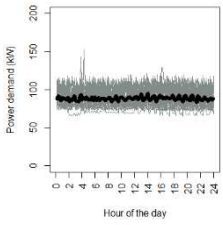
Two groups of buildings emerge from the clustering tests. For nine buildings out of fourteen – B2, B3, B5, B6, B7, B8, B10, B11 and B12 – DLP distributions are well-defined and trivially separate business days from closing days. These buildings comprise teaching and research units, two libraries and one gymnasium. DLPs are sorted into two clusters for five buildings and three clusters for the other three buildings. For two-cluster distributions, the first cluster groups most business days while the second cluster groups most closing days. A few exceptions can be noticed for these buildings with some business days in the second cluster and vice versa. These are related to specificities of building activities such as students' vacations when only part of the staff is present in the buildings and which results in a much lower electricity demand. Cases with a three-cluster distribution are more specific. The third cluster can highlight outlying DLP with higher power demand for B3 and missing data for

B12. B2 shows a slightly different behavior as it separates Saturdays (in cluster 2) and Sundays (in cluster 3). Then for B5 the overall day-type-based DLP classification is respected but the five highest DLP in terms of electric energy demand are separated from other business days and grouped in the first cluster.

The second group of building includes B1, B4 and B9 for which the analysis of clustering results with respect to the types of days fails to explain DLP distributions. Cluster 1 for B9 and cluster 2 for B1 and B4 mainly contains business days but with some closing days. Cluster 2 for B9 and cluster 3 for B1 and B4 contains significant amounts of both types of days. Then the third cluster is either grouping the highest DLPs or specific ones with a decreasing electric load demand over the day. This classification can be related to the office building type with very different activities (Table 1). Activity schedules may not be as well defined as for the first group of buildings: there are business days with low activity and power demand or, on the opposite, closing days with higher activity and then business-day-like DLPs. Therefore, it results in DLP classifications that are not only day-type-based and would require other explanatory variables to be explained.

Hence, the comparison of clustering results obtained for the fourteen buildings shows that there is a significant difference between building electricity demand depending on the building main activity. A clear trend can be identified for teaching buildings, libraries, and gym with a two or three-cluster DLP distribution: a first cluster for working days, a second one for closing days and a third cluster for outliers, although B2 separates two types of closing days. This is due to a very regular buildings operation according to the same schedule all year round with interruptions during vacation, national holidays and weekends, the overall load demand behavior. On the opposite, for office buildings, B1, B4 and B9, day-type is not a sufficient explanatory variable for electric load profiles which is probably due to a larger diversity in scheduled activities, hence in buildings occupancy.

Building	All DLP	Cluster 1	Cluster 2	Cluster 3
B1				
		7 business days	24 business days 5 closing days	31 business days 22 closing days
B2				
		62 business days	12 closing days	15 closing days
B3				
		61 business days 1 closing day	26 closing days	1 business day
B4	B6 			
		12 business days	42 business days 6 closing days	14 business days 21 closing days
B5				
		5 business days	51 business days 2 closing days	25 business days 6 closing days

B6				/
		55 business days	7 business days 27 closing days	/
B7				/
		46 business days	16 business days 27 closing days	/
B8				/
		56 business days 1 closing days	6 business days 26 closing days	/
B9				
		39 business days 1 closing days	23 business days 23 closing days	3 closing days
B10				/
		62 business days	27 closing days	/

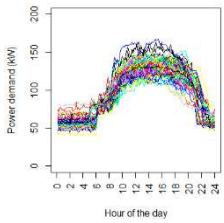
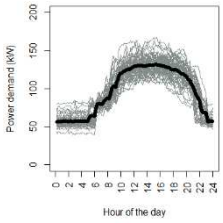
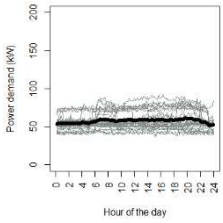
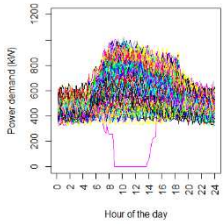
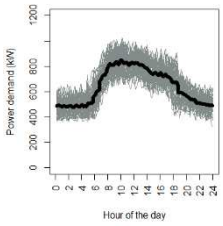
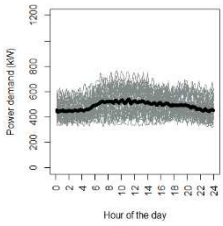
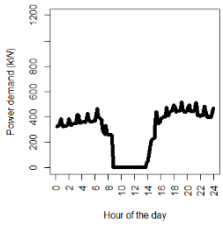
B11				/
		55 business days 1 closing days	7 business days 26 closing days	/
B12				
		49 business days	13 business days 26 closing days	1 closing day

Table 4 – Clustering results for the whole building stock using a three-month timeframe and 10-minute time-step

4.2.2 Building stock aggregated load profiles

Following the tests on all individual buildings, clustering is performed on the aggregated profiles of the building stock. A single dataset is used that considers the sum of the power demand of all the buildings for each time-step within a given timeframe.

From the previous subsection, two main groups of buildings are highlighted with respect to their main activity and their overall electricity demand pattern. However, even within a group of similar buildings there is a significant diversity of DLPs, both in terms of electricity demand characteristics (Table 5) and regarding the shapes of the profiles (Table 6). It results in a complex modelling problem when considering the buildings individually. As all fourteen buildings are located nearby each other on the same campus, an opportunity lies in considering all buildings together with their corresponding aggregated electricity demand. Thus, DLPs of the fourteen different case studies are aggregated and TS-EUCL is performed on four three-month timeframes between 05/01/2018 and 04/30/2019, with ten-minute time-

503 step. Results are presented in Table 6. In the second column are all aggregated DLPs for the
504 considered timeframe, and in the following columns the DLPs distributions are given along
505 with the number of business days and closing days for each cluster.

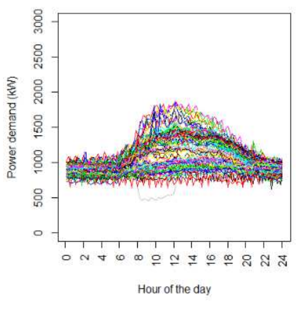
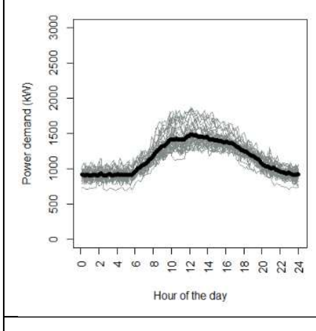
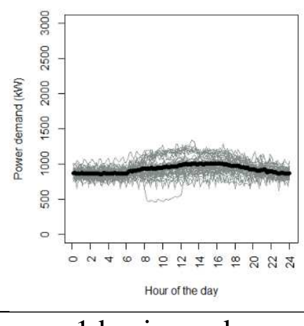
Building number	Daily mean electric power demand (kW)	Daily maximum electric power demand (kW)	Daily minimum electric power demand (kW)	Daily electric energy consumption (kWh)
B1	11.1 ±20.2	17.9 ±28.8	5.7 ±13.6	267 ±483
B2	52.7 ±17.3	74.2 ±36.7	35.1 ±9.9	1,265 ±417
B3	44.6 ±35.0	74.1 ±74.2	24.7 ±13.7	1,071 ±840
B4	23.2 ±31.3	41.0 ±53.9	13.8 ±20.0	556 ±750
B5	252.2 ±151.6	341.0 ±258.5	198.0 ±112.2	6,051 ±3639
B6	76.9 ±40.0	108.3 ±74.1	55.1 ±21.6	1,845 ±962
B7	5.3 ±5.0	9.7 ±11.0	2.3 ±1.1	128 ±120
B8	34.8 ±25.8	55.7 ±46.2	22.3 ±18.4	834 ±621
B9	8.2 ±4.3	13.7 ±9.4	5.2 ±2.2	197 ±102
B10	94.8 ±23.3	128.8 ±39.9	72.1 ±19.8	2,275 ±560
B11	75.1 ±47.9	105.8 ±77.1	47.4 ±20.0	1,801 ±1149
B12	532.5 ±197.2	717.4 ±360.6	376.1 ±124.1	12,779 ±4733

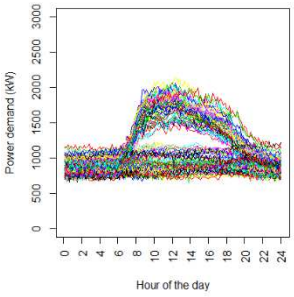
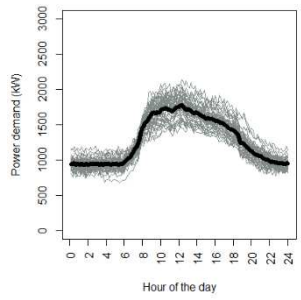
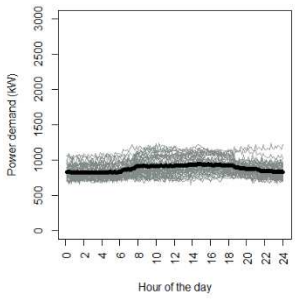
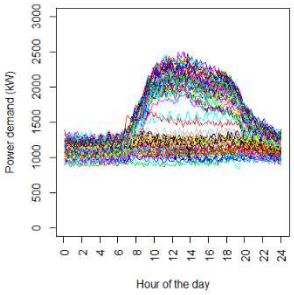
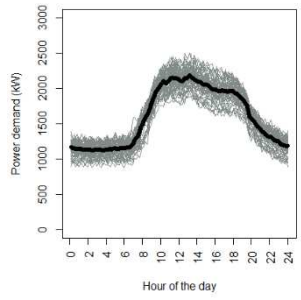
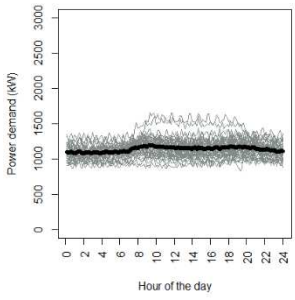
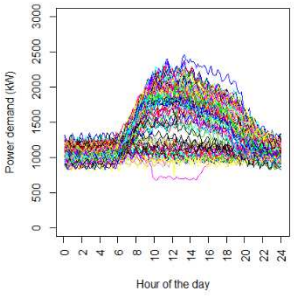
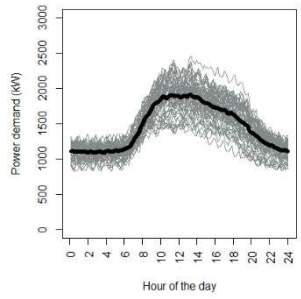
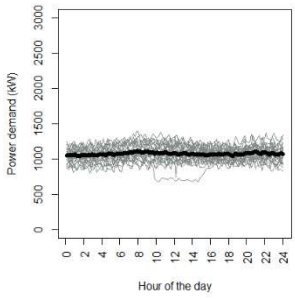
506 Table 5 – Mean, maximum, minimum electric power demand and daily electric energy
507 consumption for buildings B1 to B12 between 05/01/2018 and 04/30/2019

508 For all four timeframes, results provide a two-cluster distribution. The first cluster gathers all
509 business days including periods of students' vacations as for most buildings the activity is
510 reduced but they are not closing days. The second cluster gathers weekends, the summer
511 closing period and Christmas break when the university is closed. Only two days are found
512 out of their expected cluster: a business day with low electricity demand in cluster 2 for the
513 first timeframe and a closing day with high electricity demand found in cluster 1 for the fourth
514 timeframe. These might be related to specific events happening on the campus. Nevertheless,
515 aside from these two exceptions, DLP classifications can be easily explained using day-type
516 as an explanatory variable.

517 Thus, aggregated DLP clustering results show that the overall campus exhibits two main DLP
518 classifications depending on the type of day and the main activity on the campus. The

aggregation of occupants' behavior, appliances and activity schedules results in individual buildings' specificities that are merged in the larger load profiles. Aggregated profiles are then less impacted by building-scale events that may result in anomalous building DLPs. Furthermore, the day-type-based DLP distribution is the same as the classification observed in the previous subsection for teaching and research buildings, libraries and the gymnasium (B2, B3, B5, B6, B7, B8, B10, B11, B12). Since this group of buildings accounts for 96% for the overall mean electric power demand and daily electric energy consumption, their overall electric demand pattern is expected to be found as well in aggregated DLPs at campus-scale. Therefore, a reduction of diversity in electricity demand may be seen as an opportunity. Indeed, because of the diversity of buildings and their respective electric demand drivers, electric demand analysis and modeling is not always a straightforward task. This is particularly the case when meta-data are difficult to collect or not available as for most building in the present study. Campus-scale electric demand aggregation simplifies modeling and forecasting since it results in the emerging main day-type-based electric demand patterns for the present case study.

Observation period	All DLPs	Cluster 1	Cluster 2
2018/05/01 to 2018/07/31			
		45 business days	1 business day 45 closing days

2018/08/01 to 2018/10/31			
		43 business days	48 closing days
2018/11/01 to 2019/01/31			
		54 business days	38 closing days
2019/02/01 to 2019/04/30			
		62 business days 1 closing day	26 closing days

534 Table 6 – Clustering results for the aggregated DLP of the fourteen building case studies using TS-
535 EUCL on three-month timeframes with ten-minute time-step between 05/01/2018 and 04/30/2019

536 5 Conclusions

537 The present work reports on non-residential buildings daily electric load profile classification.
538 Fourteen buildings located on the same campus are considered. Time series of electric
539 demand with a ten-minute time-step are used as input data. A k-means algorithm is
540 implemented with three methods: clustering with feature extraction and Euclidian distance,
541 clustering of electric demand time series using Euclidian distance and Dynamic Time

Warping. The three methods are tested with different configurations of input data characteristics by varying the timeframes and time-steps and compared. We show that feature engineering-based clustering surprisingly provides very consistent results with a ten-minute time-step in spite of information loss in input data. However, it is particularly sensitive to the time-step parameter. Dynamic Time Warping is particularly sensitive to the time-step as well and provides the most accurate results with one-year timeframes. Finally, Euclidian distance clustering using electric DLP time series with three-month timeframes and ten-minute time-step outperforms all other combinations.

Conducted tests also lead to several insights related to academic building electric demand behavior. All methods are greatly affected by a seasonal effect in datasets with timeframes larger than three months which reduces classification accuracy. This seasonal effect results in the significant difference between summer business days and winter business days, as the former exhibit a much lower daily electric demand than the latter. For this reason, a particular attention should be paid to the forecasting horizon when simulating the electric demand of such buildings, for a horizon larger than three months with the present case studies. These initial results would be worth exploring in further details with clustering applications on the results of the present study to discriminate and investigate different categories of business days and closing days for different buildings.

Nevertheless, considering the whole building stock with time-series and Euclidian distance, two groups of buildings are identified. First, teaching, research, library and gymnasium buildings, which exhibit two well-defined day-type-based clusters for business days and closing days. Secondly, office buildings, which do not exhibit day-type consistent clusters. The second group of buildings shows that day-type-based trivial classification is not systematically verified. Therefore, daily load profiles classification using only electric demand data is limited and additional meta-data would be required for explanatory variables

investigation. Finally, aggregated load profiles clustering at the campus level provides two well-defined clusters distinguishing business days and closing days. Obtained results provide useful insights opportunities for non-residential buildings electric demand analysis, modeling and forecasting at different timeframes, time-steps and spatial scales.

Acknowledgment

This work was supported by the I-SITE FUTURE Initiative (reference ANR-16-IDEX-0003) in the frame of the project ANDRE. Authors would like to thank H. Hoxha and W. Wu for the development of some of the data processing algorithms, A. Bouzidi for fruitful discussions and administrative staff of University Gustave Eiffel for providing access to the data used in the present work.

Declaration of interest

None

Appendix A

Timeframe	Time-step	B12									
		FB-MAN			TS-EUCL			TS-DTW			
		C1	C2	C3	C1	C2	C3	C1	C2	C3	C4
1 month *from 01/01/2018 to 31/01/2018*	10-minute	18 business days	12 closing days	1 outlier	18 business days	12 closing days	1 outlier	18 business days	8 closing days	4 closing days	/
								18 business days	9 closing days	3 closing days	1 outlier
	30-minute	18 business days	12 closing days	1 outlier	18 business days	12 closing days	1 outlier	18 business days	13 closing days	/	/
	Hourly	18 business days	12 closing days	1 outlier	18 business days	12 closing days	1 outlier	18 business days	13 closing days	/	/
3 months *from 01/01/2018 to 31/03/2018*	10-minute	60 business days	29 closing days	1 outlier	60 business days 3 closing days	26 closing days	1 outlier	60 business days 3 closing days	27 closing days	/	/
	30-minute	14 business days 11 closing days	46 business days 19 closing days	/	60 business days 3 closing days	26 closing days	1 outlier	60 business days 3 closing days	27 closing days	/	/
	Hourly	60 business days 1 closing day	29 closing days	/	60 business days 3 closing days	26 closing days	1 outlier	60 business days 15 closing days	15 closing days	/	/
6 months *from 01/01/2018 to 30/06/2018*	10-minute	50 business days	51 business days 51 closing days	22 business days 7 closing days	69 business days 3 closing days	51 business days 58 closing days	/	69 business days 3 closing days	51 business days 58 closing days	/	/
	30-minute	72 business days 17 closing days	51 business days 41 closing days	/	69 business days 3 closing days	51 business days 58 closing days	/	62 business days 3 closing days	58 business days 58 closing days	/	/
	Hourly	72 business days	51 business days 58 closing days	/	69 business days 3 closing days	51 business days 58 closing days	/	62 business days 3 closing days	58 business days 58 closing days	/	/
1 year *from 01/01/2018 to 31/12/2018*	10-minute	134 business days	75 business days 137 closing days	15 business days 4 closing days	82 business days 3 closing days	122 business days 158 closing days	/	147 business days 6 closing days	57 business days 155 closing days	/	/
	30-minute	163 business days 6 closing days	61 business days 135 closing days	/	82 business days 3 closing days	114 business days 14 closing days	8 business days 144 closing days	160 business days 16 closing days	44 business days 145 closing days	/	/
	Hourly	148 business days	76 business days 141 closing days	/	82 business days 3 closing days	122 business days 158 closing days	/	50 business days 2 closing days	147 business days 25 closing days	7 working days 134 closing days	/

Table A.1 – Detailed results of the comparative analysis performed on B12

References

[1] International Energy Agency, Data & Statistics, (2018).

[2] International Energy Agency, Tracking Buildings – Analysis, (2020). https://www.iea.org/reports/tracking-buildings (accessed April 20, 2020).

[3] B.N. Silva, M. Khan, K. Han, Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities, Sustain. Cities Soc. 38 (2018) 697–713. doi:10.1016/j.scs.2018.01.053.

[4] D.R. Obey, Text - H.R.1 - 111th Congress (2009-2010): American Recovery and Reinvestment Act of 2009, (2009). https://www.congress.gov/bill/111th-congress/house-bill/1/text.

[5] U.S. Energy Information Administration (EIA), How many smart meters are installed in the United States, and who has them?, (2018). https://www.eia.gov/tools/faqs/faq.php?id=108&t=3.

[6] X. Liu, C. Marnay, W. Feng, N. Zhou, N. Karali, A Review of the American Recovery and Reinvestment Act Smart Grid Projects and Their Implications for China, 2017.

[7] European Commission - Joint Centre | Smart Electricity Systems and Interoperability, Smart Metering deployment in the European Union | JRC Smart Electricity Systems and Interoperability, (n.d.). https://ses.jrc.ec.europa.eu/smart-metering-deployment-european-union.

[8] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: A review, Renew. Sustain. Energy Rev. 23 (2013) 272–288. doi:10.1016/J.RSER.2013.03.004.

[9] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, P. Chatellier, Modeling and forecasting building energy consumption: A review of data-driven techniques, Sustain. Cities Soc. 48 (2019) 101533. doi:10.1016/j.scs.2019.101533.

[10] 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) 113727. doi:10.1016/j.apenergy.2019.113727.

[11] 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. doi:10.1016/j.enbuild.2014.01.002.

[12] M. Quintana, P. Arjunan, C. Miller, Islands of misfit buildings: Detecting uncharacteristic electricity use behavior using load shape clustering, Build. Simul. (2019). doi:10.13140/RG.2.2.11489.86883.

[13] 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. doi:10.1016/j.apenergy.2014.12.039.

[14] J.D. Rhodes, W.J. Cole, C.R. Upshaw, T.F. Edgar, M.E. Webber, Clustering analysis of residential electricity demand profiles, Appl. Energy. 135 (2014) 461–471. doi:10.1016/j.apenergy.2014.08.111.

[15] M. Richard, H. Fortin, A. Poulin, M. Leduc, Daily load profiles clustering : a powerful

tool for demand side management in medium-sized industries, ACEEE Summer Study Energy Effic. Ind. (2017) 160–171.

[16] 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. doi:10.1016/j.energy.2019.05.124.

[17] A. Rajabi, M. Eskandari, M. Jabbari Ghadi, S. Ghavidel, L. Li, J. Zhang, P. Siano, A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications, Energy Build. 203 (2019). doi:10.1016/j.enbuild.2019.109455.

[18] G. Chicco, R. Napoli, F. Piglione, Load pattern clustering for short-term load forecasting of anomalous days, in: 2001 IEEE Porto Power Tech Proc., 2001: pp. 217–222. doi:10.1109/PTC.2001.964745.

[19] A. Satre-Meloy, M. Diakonova, P. Grünewald, Cluster analysis and prediction of residential peak demand profiles using occupant activity data, Appl. Energy. 260 (2020). doi:10.1016/j.apenergy.2019.114246.

[20] A. Lavin, D. Klabjan, Clustering time-series energy data from smart meters, Energy Effic. 8 (2015) 681–689. doi:10.1007/s12053-014-9316-0.

[21] E.C. Bobric, G. Cartina, G. Grigoraş, Clustering techniques in load profile analysis for distribution stations, Adv. Electr. Comput. Eng. 9 (2009) 63–66. doi:10.4316/aece.2009.01011.

[22] F. Spertino, G. Chicco, A. Ciocia, S. Corgnati, P. Di Leo, D. Raimondo, Electricity consumption assessment and PV system integration in grid-connected office buildings, in: 2015 IEEE 15th Int. Conf. Environ. Electr. Eng. EEEIC 2015 - Conf. Proc., Institute of Electrical and Electronics Engineers Inc., Spertino2015, 2015: pp. 255–260. doi:10.1109/EEEIC.2015.7165548.

[23] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, Energy. 42 (2012) 68–80. doi:10.1016/j.energy.2011.12.031.

[24] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, C. Toader, Customer characterization options for improving the tariff offer, IEEE Trans. Power Syst. 18 (2003) 381–387. doi:10.1109/TPWRS.2002.807085.

[25] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S.E. Lee, C. Sekhar, K.W. Tham, k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement, Energy Build. 146 (2017) 27–37. doi:10.1016/j.enbuild.2017.03.071.

[26] Pecan Street Inc., PECAN STREET, (n.d.). https://www.pecanstreet.org/ (accessed August 20, 2009).

[27] S. Xu, E. Barbour, M.C. González, Household Segmentation by Load Shape and Daily Consumption, Proc. ACM SigKDD 2017 Conf. (2017) 1–9. doi:10.475/123.

[28] Irish Social Science Data Archive, Data from the Commission for Energy Regulation (CER) – smart metering project, (2012). http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[29] S. Haben, C. Singleton, P. Grindrod, Analysis and clustering of residential customers energy behavioral demand using smart meter data, IEEE Trans. Smart Grid. 7 (2016) 136–144. doi:10.1109/TSG.2015.2409786.

[30] R. Rashed Mohassel, A. Fung, F. Mohammadi, K. Raahemifar, A survey on Advanced Metering Infrastructure, Int. J. Electr. Power Energy Syst. 63 (2014) 473–484. doi:10.1016/j.ijepes.2014.06.025.

[31] Y. Allab, M. Pellegrino, X. Guo, E. Nefzaoui, A. Kindinis, Energy and comfort assessment in educational building - Case study in a French university campus, Energy Build. 143 (2017) 202–219. doi:10.1016/J.ENBUILD.2016.11.028.

[32] M. Bourdeau, X. Guo, E. Nefzaoui, Buildings energy consumption generation gap: A post-occupancy assessment in a case study of three higher education buildings, Energy Build. 159 (2018). doi:10.1016/j.enbuild.2017.11.062.

[33] Python, Python, (2017). https://www.python.org/.

[34] NumPy, NumPy, (n.d.). https://numpy.org/ (accessed April 24, 2020).

[35] The R Foundation, The R Project for Statistical Computing, (n.d.). https://www.r-project.org/.

[36] M. Charad, N. Ghazzali, V. Boiteau, A. Niknafs, NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set, J. Stat. Softw. 61 (2014).

[37] P. Montero Manso, J.A. Vilar, TSclust: Time Series Clustering Utilities, 2017.

[38] H. Kang, The prevention and handling of the missing data, Korean J. Anesthesiol. 64 (2013) 402–406. doi:10.4097/kjae.2013.64.5.402.

[39] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renew. Sustain. Energy Rev. 82 (2018) 1027–1047. doi:10.1016/J.RSER.2017.09.108.

[40] T. Teeraratkul, D. O'Neill, S. Lall, Shape-Based Approach to Household Electric Load Curve Clustering and Prediction, IEEE Trans. Smart Grid. 9 (2018) 5196–5206. doi:10.1109/TSG.2017.2683461.

[41] C. Miller, Z. Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renew. Sustain. Energy Rev. 81 (2018) 1365–1377. doi:10.1016/j.rser.2017.05.124.

[42] T. Warren Liao, Clustering of time series data - A survey, Pattern Recognit. 38 (2005) 1857–1874. doi:10.1016/j.patcog.2005.01.025.

[43] T. Räsänen, M. Kolehmainen, Feature-based clustering for electricity use time series data, in: Int. Conf. Adapt. Nat. Comput. Algorithms, 2009: pp. 401–412. doi:10.1007/978-3-642-04921-7_41.

[44] C. Fan, F. Xiao, S. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Appl. Energy. 127 (2014) 1–10. doi:10.1016/J.APENERGY.2014.04.016.

[45] D.J. Bemdt, J. Clifford, Using DynamicTime Warping to Find Patterns in Time Series, KDD-94 Work. Knowl. Discov. Databases. 398 (1994) 359–370.

[46] K.A. Choksi, S. Jain, N.M. Pindoriya, Feature based clustering technique for investigation of domestic load profiles and probabilistic variation assessment: Smart meter dataset, Sustain. Energy, Grids Networks. 22 (2020) 100346. doi:10.1016/j.segan.2020.100346.