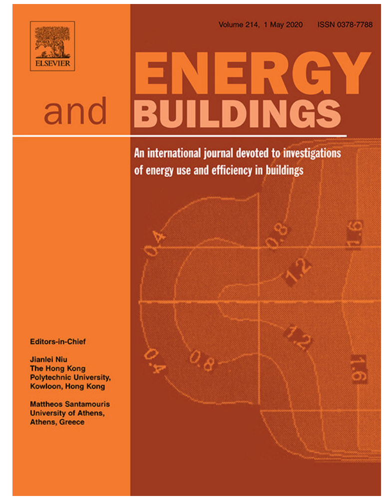
Journal Pre-proofs



A reduced-dimension feature extraction method to represent retail store electricity profiles

Ramon Granell, Colin J. Axon, Maria Kolokotroni, David C.H. Wallom

PII: S0378-7788(22)00679-X
DOI: https://doi.org/10.1016/j.enbuild.2022.112508
Reference: ENB 112508

To appear in: *Energy & Buildings*

Received Date: 19 May 2022
Accepted Date: 20 September 2022

Please cite this article as: R. Granell, C.J. Axon, M. Kolokotroni, D.C.H. Wallom, A reduced-dimension feature extraction method to represent retail store electricity profiles, *Energy & Buildings* (2022), doi: https://doi.org/10.1016/j.enbuild.2022.112508

A reduced-dimension feature extraction method to represent retail store electricity profiles

Ramon Granell[a,b,1], Colin J. Axon[a], Maria Kolokotroni[a], David C.H. Wallom[b]

[a]*Institute of Energy Futures, Brunel University London, Uxbridge, London UB8 3PH, United Kingdom.*
[b]*Oxford e-Research Centre, Department of Engineering Science, University of Oxford, 7 Keble Road, Oxford OX1 3QG, United Kingdom.*

Abstract

Characterising the inter-seasonal energy performance of buildings is a useful tool for a business to understand what is normal for its portfolio of premises and to detect anomalous patterns of energy demand. When adding a new building to the portfolio, it will be useful to predict what will be the likely energy use as part of on-going monitoring of the site. For a large portfolio of buildings with, say, half-hourly energy use measurements (48 dimensions), analysis and prediction will require machine learning tools. Even so, it is advantageous to minimise the amount of data and number of dimensions and features required to find useful patterns in the measurement stream. Our aim is to devise a reduced feature set that can generate a statistically reasonable representation of daily electricity load profiles of retail stores and small supermarkets. We then test if our method is sufficiently accurate to predict and cluster measured patterns of demand. We propose an automatic method to extract features such as times and average demands from electricity load profiles. We used four regression models for prediction and six clustering methods to compare with the results obtained using all of the readings in the load profile. We found that the reduced feature set gave a good representation of the load profile, with only small prediction and clustering errors. The results are robust as prediction is supervised learning and clustering is unsupervised. This simplified feature set is a concise way to represent profiles without using small variances of the demand that do not add useful information to the overall picture. As modern sensor systems increase the volume, availability, and immediacy of data, using reduced dimensional

Email address: `ramon.granell@oerc.ox.ac.uk` (Ramon Granell)
[1]Corresponding author.

datasets will be key to extracting useful information from high-resolution data streams.

Keywords: clustering, electricity demand, commercial, prediction, machine learning, supermarket

---

Abbreviations

ANN Artificial neural networks

ED Euclidean distance

EDLP Electricity daily load profile

kNNR k-nearest neighbours regression

ML Machine learning

NP Normalised percentage difference with respect to the original EDLP

OLS Ordinary least of squares

SE Supermarkets using only electricity

SEG Supermarkets using electricity and gas

SVR Support vector regression

Symbols

e_i electricity consumed (kWh) between the $(i-1)$-th and i-th time interval

k number of EDLPs used for the prediction

p number of previous years used to predict the EDLP

s_0 off-peak time period in the EDLP

s_1 time period of the off-peak to peak transition time in the EDLP

s_2 peak time period in the EDLP

s_3 time period of the peak to off-peak transition time in the EDLP

t_0 first time interval of the EDLP where the slope of the off-peak/peak transition starts

t_1 first time interval of the EDLP where the main peak stabilises

t_2 first time interval of the EDLP where the peak starts to decrease

t_3 first time interval of the EDLP where the non-peak behaviour stabilises after the peak

$\vec{t}$ t_0, t_1, t_2 and t_4

y year used to compute the EDLP

D number of time intervals of the EDLP

B set of supermarket building characteristics used to predict the EDLP

L_s EDLP of the supermarket s

S, S' sets of new and existing supermarkets respectively

2-feat $\mu(s_0)$, $\mu(s_2)$, $m(s_1)$, $m(s_3)$ and $\vec{t}$

4-feat $\mu(s_0)$, $\mu(s_2)$, $m(s_1)$ and $m(s_3)$

8-feat $\mu(s_0)$, $\mu(s_2)$, $m(s_1)$, $m(s_3)$ and $\vec{t}$

$\mu(s_i)$ mean of the energy values that are in s_i

$m(s_i)$ slope of the line that crosses the energy values that are in s_i

1. Introduction

The aim of reducing greenhouse gas emissions is shared by most countries [1, 2] with the UK aiming at greenhouse neutrality by 2050 [3]. As energy use in buildings across the EU accounts for more than 30% of final energy demand [4], cutting and time-shifting energy demand of all types of buildings (residential, commercial, services and industrial) are needed to achieve these targets. Residential buildings have received much attention [5, 6, 7], whilst commercial and industrial buildings less so due to the lack of open data-sets and their heterogeneity [8, 9].

The total energy demand and the temporal profile are useful performance indicators for buildings estate management, investment decisions, site acquisitions, and improvement programmes. Knowing the expected demand of a store establishes a baseline for: 1) planning annual energy budgets for the portfolio of stores, 2) negotiating energy supply contracts, and 3) detecting stores with abnormal or anomalous usage. Inevitably there are differences between stores, with the key being understanding the variability and what is acceptable usage for any store. Grouping the stores based on common demand patterns reveals existing distinct behaviours in the store portfolio [10, 11, 12, 13]. This informs which measures might be more effective or cost-efficient for each group, and identifies stores whose demand patterns do not match any of the discovered groups (anomalous behaviour).

In general, clustering techniques are unsupervised machine learning algorithms that divide data-sets into groups (clusters) without a priori information [14, 11]. Both prediction and clustering of energy demand is commonly performed over electricity daily profiles (EDLP), which are a concise, informative, and intuitive way to represent, analyse and visualise the electricity demand of any source [11, 12]. EDLPs are data representations for which the electricity demand during a day is computed with a temporal granularity, D. This temporal resolution indicates the number of points (demand values) that formed the profile, *e.g.* if $D = 24$ each demand value is the hourly demand. A disadvantage of using full EDLPs is the so-called "curse of dimensionality" [15], meaning that some machine learning (ML) algorithms can have temporal and memory issues when working with high-dimensional data-sets. By the same token, too few data points for training the algorithms risk over-fitting the model.

We propose representing EDLPs using only a small set of characteristic features (dimensional reduction) that are automatically extracted from the profile instead of using the D-dimensional daily profile. We separately investigate both predicting and clustering the EDLPs using only the extracted features. New supermarkets profiles are predicted using the historical demand of other stores. Four different ML regression algorithms are implemented to predict EDLPs over a data-set of 213 supermarkets during a six year period. Experiments in clustering EDLPs are performed using six different algorithms over the supermarket data-set and a data-set of 641 retail stores, independently. Both data-sets are real data obtained from smart meters. Based on the proposed extracted features and these two ML problems, the questions that we try to answer are:

- How accurately can D-dimensional EDLPs be represented using a small set of features?
- Using only this set of features is it possible to predict future EDLPs of new stores with different ML methods, as accurately as when using the whole EDLP?
- Using only this set of feature is it possible to cluster the electricity demand as accurately as when using the whole EDLP?
- Is it possible to extend and generalise this representation over other commercial data-sets that have other temporal resolution?

The paper is structured in the following way. We review the literature for ML related to energy analytics is in Section 2. In Section 3 we explain the pre-processing of the real-world data-sets and the computational experiments. The results obtained for these experiments and discussion about them are in Section 4. Finally, we draw conclusions and propose future work in Section 5.

2. Literature review

Predicting electricity demand of buildings (regardless of type) can be divided into two basic approaches: model-driven and data-based. The model-driven approach uses sophisticated high-resolution engineering methods based on the thermal, energy, and architectural features of the building to simulate its energy demand. In data-driven approaches, the energy performance of the building is directly modelled with numerical and statistical methods. There are extensive reviews on methods to predict, model and benchmark energy use in buildings [16, 17, 18, 19, 20, 21]. A possible classification of the data-driven techniques used to predict energy demand [20, 21] is: 1) conventional statistical techniques, 2) classification-based models, 3) support vector regression (SVR) model, 4) artificial neural networks (ANN), 5) genetic algorithms, 6) grey models, 7) fuzzy model and 8) other models (*e.g.* case-based reasoning). Our study exploits the first four classes of techniques, and we focus our review on the prediction of demand in commercial buildings and supermarkets.

Conventional statistical techniques include change-point algorithms and linear regression models such as autoregressive models and ordinary least squares (OLS). Autoregressive models have been used to predict short-term heat load for a single building [22] and, in combination with ANN, used to predict the annual electricity consumption of 787 education facilities in South Korea over a period of seven years [23]. Schrock and Clarige [24] used a change-point algorithm and a year of 15-min electricity readings of one grocery store to predict hourly and daily consumption. Linear regression has been applied to the prediction of 1-h heat load profiles of 116 buildings (health, education, business, and hotels) over three years [25]. The same linear models have also been used on data from 215 UK large supermarkets to estimate the total annual electricity demand [26], and by [27] to estimate annual energy-use intensity for UK 30 supermarkets using building features such as floor area and building age and the number of customers. In the context of climate change adaption [28] exploited temperature and humidity values to predict weekly electricity and gas demand for a single supermarket for the period 2030-2059 using multiple linear regression analysis.

Classification-based models include algorithms that were extended to perform regression. The k-nearest neighbour regression (kNNR) algorithm was used to forecast the next day consumption of 6,000 domestic Irish buildings in [29], and for the hourly air conditioning load of an office building in China [30]. Random forest (set of decision trees) and ANN (separately) were used to predict hourly HVAC loads of a Spanish hotel [31] over a period of 15 months. Similarly, decision trees, ANN and linear regression are compared to predict weekly electricity consumption of 1200 dwellings during the winter and summer of one year [32].

ANN has been used to predict the energy demand in 17 studies from 1996-2015 [19]. Short-term electricity demand of a commercial building complex using 15-min resolution data was predicted using ANN in [33]. Daily diurnal cooling load is forecasted for three university buildings with ANN in [34], using data recorded over two years. Both ANN and SVR were compared when predicting hourly cooling load in an office building in China [35] and hourly energy consumption of an office building in Shanghai [36]. Electricity consumed by the HVAC and refrigeration systems of one supermarket is predicted using ANN [37]. ANN, Gaussian process regression, linear regression and dynamic mode decomposition are compared in the prediction of 1-h weekday profiles of a commercial building [38]. Lastly, deep learning models (large neural-networks) have been also explored for this problem, however they need large data-sets to estimate the model parameters. For example, Hafeez *et al.* [39] used deep learning for short-term load forecasting over three power-grids with hourly resolution. A deep learning network and a genetic algorithm were combined to predict the 1-h daily profile in an office building over one year [40]. This work applies the clustering of daily weather profile before predicting the demand.

Support Vector regression (SVR) models were used by Dong *et al.* [41] to predict monthly energy consumption of four commercial buildings in Singapore. Models based on SVR have also been used to predict the energy load (hours to days) of a French residential building [42]. SVR and six other techniques was also investigated by [43] to predict next-hour residential building electricity consumption for three houses. Jain *et al.* [44] examine the impact of temporal (*e.g.* daily, hourly, 10 min intervals) and spatial (*e.g.* , whole building, by floor, by unit) granularity to short-term prediction. Experiments were performed using SVR over data from a multi-family residential building in the USA. Granell *et al.* [13] compared four techniques, kNNR, ordinary least of squares linear regression, ANN, and SVR in predicting whole EDLPs of new supermarkets using data from a portfolio of 213 UK supermarkets with readings spanning six years.

From this range of techniques we can conclude that there is no consensus about the superiority of a specific technique. Studies that compare several techniques usually report marginally differences in the prediction results *e.g.* [30, 32, 13, 43], or contradictory results *e.g.* ANN over-performs SVR [36] and vice-versa [35]. These results support our selection of four different types of predictors to address our problem. In addition, our prediction work addresses some of the areas that have received less attention. First, retail is clearly under-represented in the literature. For example, according to reviews by Chung [16] and Li *et al.* [21] only 22% and 33%, respectively, of investigations were about consumption in commercial buildings, and fewer still in other studies [17, 18]. Particularly notable is the severe lack of work in the literature on predicting energy use by supermarkets. There are differences between patterns of energy demand in commercial and retail premises, but also similarities in niche sectors [8]. Secondly, the prediction of daily profiles [39, 25, 13, 40] is not common, most of the long-term prediction studies use weekly, monthly, or annual demand. Thirdly, prediction experiments using retail data-sets with a size that can be considered representative (hundreds of buildings) are also infrequent. Finally, predicting the future demand of new buildings for a long period of time (more than three or four years) is a highly unusual approach; most studies predict the future demand for the study building and they usually do not use several years of continuous data.

Reviews of clustering methods applied over electrical data can be found in [11, 12, 10]. Most studies have used residential data-sets, but some work clustering electricity profiles of commercial and industrial customers has been completed. For example, 292 Greek industrial and service customers are clustered using a two-stage ML algorithm [45]. Wavelet decomposition was used [46] to select significant features to describe the hourly load profiles of 9,092 Danish industrial and commercial loads for two-week data. Later, they applied clustering using the k-means algorithms over these features. In [47] they investigate several clustering techniques such as k-means and hierarchical algorithms to cluster 234 non-residential customers, and a data set of 1,877 UK business from the entertainment sector was used to perform clustering with a Dirichlet process mixture model [48].

A recent review of dimensional reduction techniques appears in [10]. Dimensional reduction has been attempted for electricity demand modelling and clustering [46], and for symbolic aggregate approximation with hierarchical clustering [49]. Representing the data with principal component analysis, the curvilinear component analysis, and the Sammon map are investigated by [47]. The effect of the time resolution when clustering domestic EDLPs [50] was investigated by averaging over regular intervals instead of extracting key features based on the specific shape of the retail EDLP as we do here. Residential demand profiles have been characterised and clustered with a set of five points that match the peaks [51].

3. Methods

First we describe the data-sets used to perform the experiments. Secondly, the features to represent the EDLP and methods to extract them are explained. Thirdly, we describe the prediction algorithms, evaluators, and experiments. Finally, clustering algorithms and evaluators are defined.

3.1. The data-sets

Two data-sets are used to perform the experiments. The first comprises 1-h resolution electricity meter readings (kWh) from 213 UK supermarkets of the same chain for the period 2012–17. The detail of the meta-data features available of each supermarket are described elsewhere [13], but are summarised as: floor area subdivided into eight categories (*e.g.* chilled, produce, storage), geographical location, daily average external temperature, and electricity consumption. There are 129 supermarkets that use electricity and gas (SEG) and 84 supermarkets that use only electricity (SE). The second data-set comprises 663 UK retail stores (single company) with electricity meter readings at 0.5-h resolution acquired between April 2013 and October 2014. In this case, the only meta-data fields are the address and outlet type category that summarise the location of the store (*e.g.* arterial route, high street retail park, shopping centre).

For both data-sets an analytic filtering pre-process removes anomalous readings with zero or negative values, accounting for less than 0.8% of the data. In addition, stores with less than the equivalent of half a month of data (360 and 720 readings for the supermarket and retail store data-sets respectively) are removed: For the retail store data-set this was 22 shops leaving 641 stores for analysis, whilst for the supermarkets it varied from year-to-year [13].

3.2. Features extraction to represent the EDLP of supermarkets

Like most retailers, the supermarkets have a fixed daily schedule: they usually open in the morning to close later in the evening [52]. Based on these schedules, the electricity consumption patterns are quite similar to each other with a typical inverted-U shape. Figure 1 shows the daily profiles, for different seasons, of four different supermarkets and retail stores from our data-sets. These eight EDLPs show similar patterns characterising the peak and off-peak periods, however, they exhibit variability during these periods. Energy demand by supermarkets are greater than that of retail stores.

Based on these behaviours we can define four time periods in which important changes occur (Figure 2):

- t_0 indicates the first time interval of the EDLP where the slope of the off-peak/peak transition starts.
- t_1 is the first time interval of the EDLP the main peak stabilises.
- t_2 is the first time interval of the EDLP the peak starts to decrease.
- t_3 is the first time interval of the EDLP where the non-peak behaviour stabilises after the peak.

These periods follow the conditions that $t_i \in [0, D-1], 0 \leq i \leq 3$ and $t_i < t_{i+1}, 0 \leq i \leq 2$. In the example given in Figure 2 their value are: $t_0 = 6$, $t_1 = 9$, $t_2 = 15$ and $t_3 = 21$, corresponding to 6.00am, 9.00am, 3.00pm and

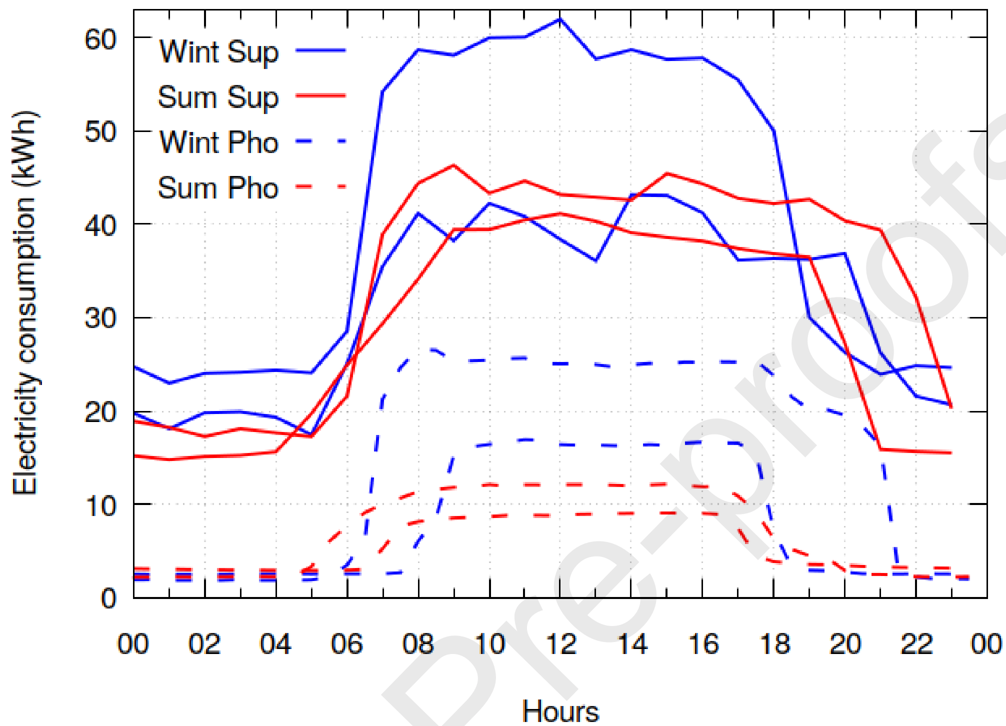


Figure 1: Example Winter and Summer daily profiles of four different supermarkets (Sup) and retail stores (Pho).

9.00pm, respectively. By defining the vector grouping the four time features as $\vec{t} = \{t_0, t_1, t_2, t_3\}$, we can divide the EDLP into four intervals using:

off-peak time period in which the supermarket is closed and the demand is a stable baseload of refrigeration, as HVAC and lighting should be switched-off or to minimum power. Formally, it is $s_0 = [0, t_0 - 1[\cup [t_3, D-1]$ *e.g.* horizontal green lines in Figure 2.

off-peak to peak transition short period occurring a little before the store is opened to customers when the HVAC, lighting, and other services ramp to their peak values. Formally, it is $s_1 = [t_0 - 1, t_1]$ *e.g.* horizontal yellow line in Figure 2.

peak period in which the demand is constantly high as the supermarket is open. The appliance power consumption is usually stable, but short-term variability may occur (see EDLPs of Figure 1). Formally, it is $s_2 = [t_1, t_2 - 1]$ *e.g.* horizontal pink line in Figure 2.

peak to off-peak transition short period following the closure of the store to customers, but staff may still be present. Modern appliances should not have a very long temporal lag for reducing their demand when they are switched-off. Formally, it is $s_3 = [t_2 - 1, t_3]$ *e.g.* horizontal grey line in Figure 2.

Given any interval of time $s = [t, t']$ with $t' > t$, we define two generic operators: 1) $\mu(s)$ as the mean of the energy values from time t to t', *i.e.* $\mu(s) = \sum_{i=t}^{t'} e_i/(t' - t + 1)$ and 2) $m(s)$ is the slope of the line that crosses the points (t, e_t) and $(t', e_{t'})$, *i.e.* $m(s) = (e_{t'} - e_t)/(t' - t)$.

We can describe the profile using eight features: the four time periods of the events $(\vec{t})$, consumption of the off-peak and peak periods ($\mu(s_0)$ and $\mu(s_2)$), and the slopes of the transitions ($m(s_1)$ and $m(s_3)$). The demand values of $\mu(s_0)$ and $\mu(s_2)$ are the average during all the values of the off-peak and peak respectively, and they are a linear approximation of the demand during these time intervals. Values of $m(s_1)$ is the rate of demand increasing by hour when moving from off-peak to peak period (this value is always positive as demand increases during this period.). The value of $m(s_3)$ is always negative as the demand decreases during the peak/off-peak transition interval. Given these eight features, the estimated profile $\vec{e'} = \{e'_0, \ldots, e'_{D-1}\}$ can be reconstructed using Euclidean geometry:

- Off-peak values are equal to $\mu(s_0)$:, $e'_i = \mu(s_0), 0 \leq i < t_0$ and $t_3 \leq i < D$

- Values of the off-peak/peak transition are computed with the linear equation $y = x * m(s_1) + b$ where independent term b is computed by substituting the equation with the data point $(t_0 - 1, \mu(s_0))$: $e'_i = i * m(s_1) + b, t_0 \leq i < t_1$.

- Peak values are equal to $\mu(s_2)$: $e'_i = \mu(s_2), t_1 \leq i < t_2$.

- Values of the peak/off-peak transition are calculated with the linear equation $y = x * m(s_3) + b'$ where term b' is computed by substituting equation with the data point $(t_2 - 1, \mu(s_2)$: $e'_i = i * m(s_3) + b', t_2 \leq i < t_3$.

As an example, Figure 2 shows a reconstructed profile (red lines) obtained using the eight selected features and a real profile (black line) of a supermarket EDLP. The off-peak demand is estimated well, likewise the central part of the peak demand. However, the beginning of the peak demand is underestimated and the end is overestimated. The discrepancy (error) between the reconstructed profile ($\vec{e'} = \{e'_0, \ldots, e'_{D-1}\}$) and the real values of the profile ($\vec{e} = \{e_0, \ldots, e_{D-1}\}$) is quantified using evaluators:

Euclidean Distance (ED) in which discrepancies between the EDLPs absolute values are accumulated (in kWh),

$$\sqrt{\sum_{i=0}^{D-1}(e_i - e'_i)^2} \tag{1}$$

Normalised Percentage (NP) difference with respect to the original EDLP (NP) computes the relative distance considering the proportion of the error with respect to the total consumption of the original profile,

$$\frac{100 * \sum_{i=0}^{D-1} |e_i - e'_i|}{\sum_{i=0}^{D-1} e_i} \tag{2}$$

The ED and the NP between the modelled and real EDLPs of Figure 2 are 15.8 kWh and 3.9% respectively. The evaluators $\overline{ED}$ and $\overline{NP}$ are extended over the whole data-set using the average ED and NP respectively for all stores.

As the whole feature set can be obtained directly with the time period vector $\vec{t}$, they can be automatically computed searching using the objective function to minimise the error:

$$\hat{\vec{t}} = \arg\min_{\vec{t}}(\mathrm{Ev}(\vec{e}, \vec{e'}_{\vec{t}})) \tag{3}$$

where $\vec{e'}_{\vec{t}}$ is the reconstructed profile using $\vec{t}$, and Ev is an evaluator computed over the two EDLPs. For evaluator Ev, we use the ED. A brute-force search method in which all possible values of $\vec{t}$ are explored to find the optimal solution $\hat{\vec{t}}$ as it is restricted search. For the example (Figure 2) the set of features obtained using this objective-function method are $\vec{t} = (6, 9, 15, 21)$, $\mu(s_0) = 32.0$ kWh, $\mu(s_2) = 100.0$ kWh, $m(s_1) = 16.2$ kWh/h and $m(s_3) = -9.2$ kWh/h. The utility of this approach needs to be demonstrated for problems such as prediction and clustering.

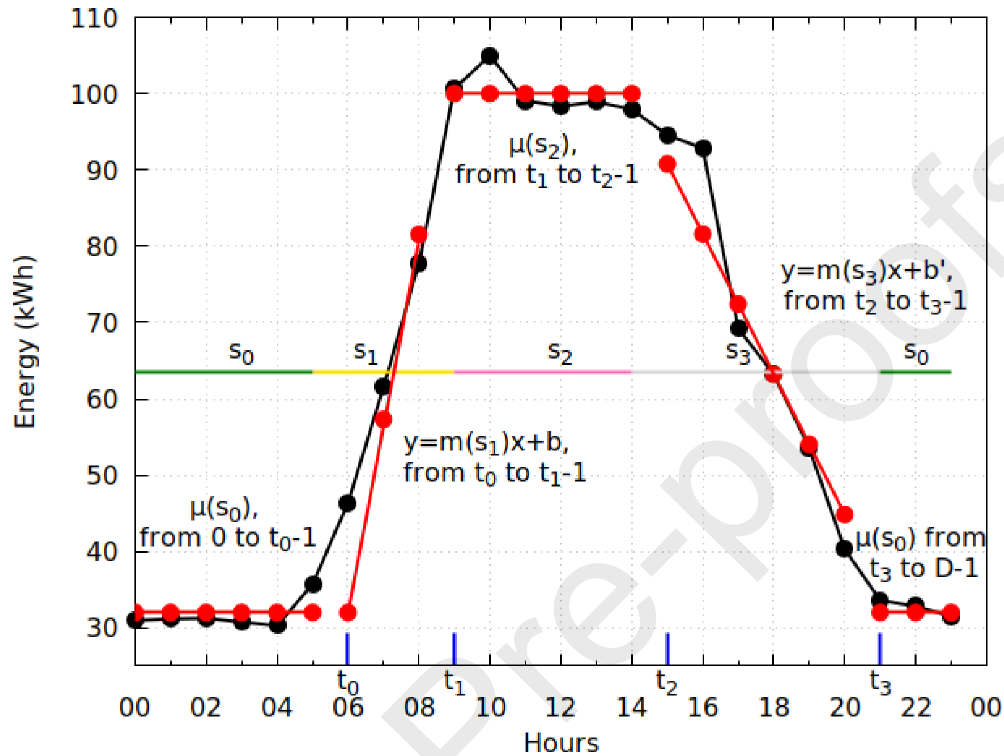


Figure 2: Modelled profile based on the eight proposed features (red line) and real profile (black line).

3.3. Computational prediction experiments

Experiments are performed using only the extracted features to predict electricity demand of new supermarkets. The EDLP of a new supermarket $L_s = e_0, \ldots, e_{D-1}$ for a year y is predicted based on historical profiles of existing supermarkets S' and the supermarket building characteristics B. L_s is the EDLP of the new supermarket s, e_i is the electricity consumed (kWh) between the $(i-1)$-th and i-th time interval, D is the number of intervals, S and S' are the set of new and existing historical supermarkets, respectively ($S \cap S' = \emptyset$). The set of store characteristics B is the set of available information about the supermarket building such as the floor area divided by usage and the supermarket geographical location. Therefore, we train a regression ML algorithm with all the supermarkets of S' where the independent variables (input) are the store characteristics B and the dependent variable to predict (output) is e_i, $0 \leq i < D$.

As we do not know which store characteristics B use nor how many stores k to select to train the ML model, the best combination of (k, B) is searched using Equation 4.

$$(\hat{k}, \hat{B}) = \arg\min_{k,B} \sum_{s \in S} \mathrm{Ev}(L_s, L_s(k, B))) \quad (4)$$

where S is the set of new supermarkets, L_s is the real EDLP of supermarket s, $L_s(k, B)$ is the predicted energy profile when using parameters (k, B) and $\mathrm{Ev}(L_s, L'_s(k, B))$ is the evaluator that measures the error between the predicted and real profile. As Ev we use the average Euclidean distance over all the real and predicted stores:

$$\overline{ED} = \frac{\sum_{s \in S} ED_s}{|S|} \quad (5)$$

where ED_s is the ED computed over the real and predicted EDLPs of the supermarket s.

Four different ML algorithms are investigated:kNNR [14], ANN [14], SVR [53], and OLS [54].

We only use the extracted features to represent the ELDP *i.e.* these features are predicted using as input the store characteristics B instead of predicting the whole profile. The diagram of Figure 3 illustrates the steps of the experimental set-up:

1. The eight features of each supermarket ($\vec{t}$, $\mu(s_0)$, $\mu(s_2)$, $m(s_1)$ and $m(s_3)$) are computed.
2. These features are predicted independently for each supermarket s' using the regression model using as input the store features (B'_s). That is, for each supermarket s', the eight features of the EDLP of year y are predicted with the regression algorithm. This ML model is trained with the features extracted of the EDLP computed in previous years $y - t$ of the stores of the set $S - \{s'\}$.
3. The profile of the predicted store is reconstructed with the eight predicted features of the store ($\vec{t'}$, $\mu'(s_0)$, $\mu'(s_2)$, $m'(s_1)$ and $m'(s_3)$). The evaluators are computed between this reconstructed profile and the original profile of the test supermarket (s').
4. Parameter search (k, B) is performed and final error is computed over the best parameter combination $(\hat{k}, \hat{B})$ that minimizes Equation 4.

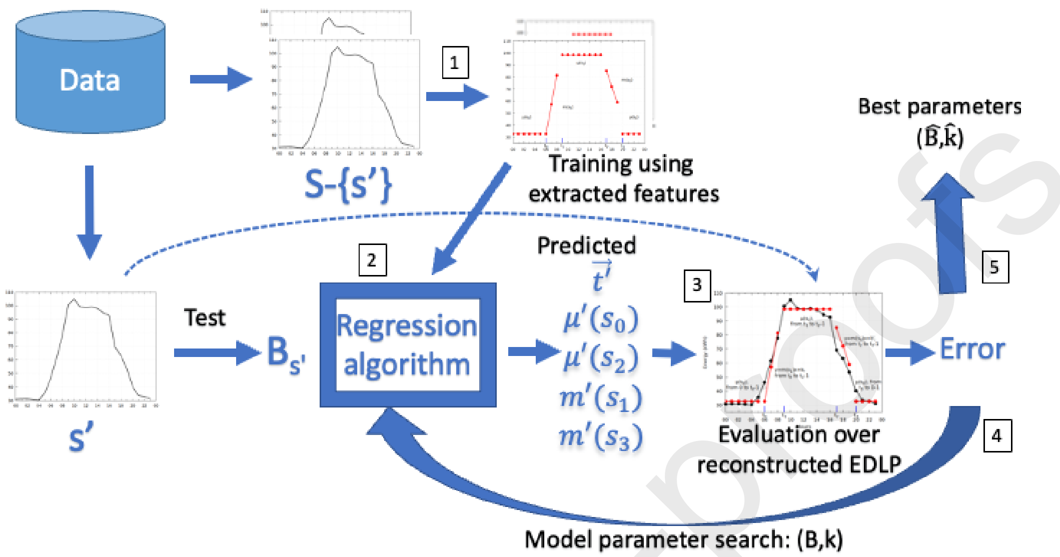


Figure 3: Logical flow of the prediction experiments using the features to represent the profiles.

The two essential points of this experimental set-up are 1) the ML algorithm predicts the summarised features of the profile, and 2) the evaluation is performed comparing the reconstructed profile using the predicted features with the real profile to predict (not with the reconstructed profile over the real features). Due to the second point it is feasible to compare the results obtained with these experiments with the results obtained when predicting the whole profile. As the values of $\vec{t}$ are integer numbers, the closest integer is selected to the value returned by the regression model.

Fewer than 30 supermarkets are opened each year and we assume that each is opened within year y. The historical EDLPs of the other $|S|-1$ supermarkets are used to predict the EDLPs of the new ones, improving the robustness of the experiments. The leaving-one-out technique [14] uses all the data points — except the one being estimated — as predictors (repeated $|S|$ times) to compute the EDLP of the new one for year y.

Experiments are carried out separately over EDLPs of the supermarkets computed for 2013–2017, seasons (Winter, Summer and Spring/Autumn), and SE/SEG sets [13]. We employ the brute-force approach (Equation 4) to search all parameter combinations $(\hat{k}, \hat{B})$. The maximum number of combinations, for each season is $(2^{|F|}-1)*(|S|-1) = (2^{|11|}-1)*(129-1) = 262,016$.

For ANN and SVR (temporally more complex) we used stepwise regression [14] with the whole feature set B (using all the supermarkets, $k = |S|$). For the ANN, we use a logistic function as an activation function, over a two internal layer net, *i.e.* the configuration of the network is $|B|$-4-2-1, where $|B|$ is the number of features. The function neuralnet of the R language [55] is used with the default parameters, *i.e.* the resilient backpropagation algorithm with 10^5 maximum steps for the net training. For SVR, we used a radial basis kernel function to model the non-linearly. The function svm of the R language [56] is used with default parameters. The parameters of the ML methods are the same that were used in [13] to enable comparison with previous work. Both R scripts were invoked for each one of the computing experiment from the generic C++ code.

3.4. Clustering experiments

Clustering experiments group all the available EDLPs computed during a specific year for each data-set independently. The result depend on both the algorithm and the way the data is represented. Our aim is to compare clustering results—not algorithm performance—with the two data representations. Thus we selected two types of clustering algorithm: partitioning and agglomerative hierarchical. The partitioning algorithm we chose was k-means [47, 11, 45, 57, 58]. For the agglomerative hierarchical algorithm [47, 11, 45, 57] there is more choice depending on the criterion used to compute the distance to merge the clusters: Single link algorithm, Complete link algorithm, Unweighted pair group method average algorithm (UPGMA), Unweighted pair group method centroid algorithm (UPGMC), Weighted pair group method centroid algorithm (WPGMC) and Ward or minimum variance algorithm (WARD).

We selected six evaluators [59] to asses the clustering results: the clustering dispersion indicator (CDI), Davies-Bouldin index (DBI), modified Dunn Index (MDI), mean index adequacy (MIA), scatter index (SI), and variance ratio criterion (VRC). These evaluators are based on the similarity of the data elements within each cluster, and the difference among elements of the other clusters.

The clustering is performed using directly three sets of features:

8 features (8-feat): $\mu(s_0)$, $\mu(s_2)$, $m(s_1)$, $m(s_3)$ and $\vec{t}$.

4 features (4-feat): $\mu(s_0)$, $\mu(s_2)$, $m(s_1)$ and $m(s_3)$.

2 features (2-feat): $m(s_1)$ and $m(s_3)$.

However, we decided to evaluate directly over the whole profiles. The reason for this is that the output of clustering is the grouping in which all the data-points (in our case EDLP) can be distributed based on the ML algorithm. As all the evaluators use the intra-point distance, we consider that the fairest way to compare the quality of the obtained grouping is to compare over the same set of points. Clustering results using the eight features are compared with respect to the clustering obtained using the whole EDLP. For the k-means algorithm 100 repetitions with different random initialisation were performed and the evaluations are averaged. The number of clusters (input parameter of the algorithm) is 2–10 exploring all the values. All the software was coded in C++.

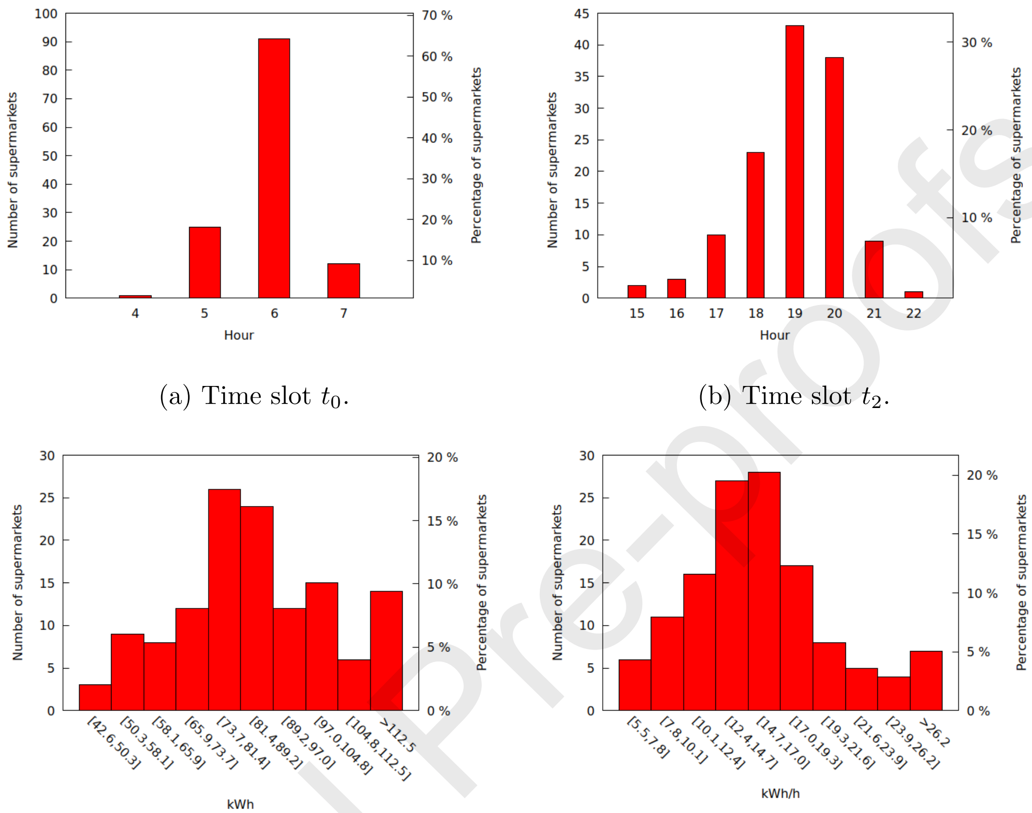
4. Results and Discussion

We have performed a large number of computational experiments. For clarity, we discuss separately the results obtained for: 1) representing the EDLPs using the eight features, 2) the prediction experiments and 3) the clustering experiments. Prediction experiments were not performed using the retail stores data-set as there was only one year of data.

4.1. Representing supermarket EDLPs with the selected features

An example of the features for the Winter 2017 profile of a SEG supermarket is in Section 3.2. Histograms of Figure 4 show the range of values for the features t_0, t_2, $m(s_1)$ and $\mu(s_2)$ extracted from the Winter 2017 profiles of all the 129 SEG supermarkets.

For the periods t_0 (Figure 4a) and t_1, there are only four different hours in which they occur, and one of the hours is much frequent than the others: 6am (70.5% of supermarkets) and 8am (50.4%) for t_0 and t_1 respectively. The period t_3 also has one value more frequent than the others (9pm, 55.0%), however there are eight different values for the t_2 (Figure 4b). The histograms exhibit little variability of values and the distribution is Gaussian. However, the most important insight is the variability in which the peak and off-peak can begin and end. This shows that using a fixed time for these moments is an over-simplification that does not properly represent the real pattern of the demand. In addition, the range of values for these time slots is restricted, indicating common patterns for the supermarkets. Intervals for the mean



(a) Time slot t_0.

(b) Time slot t_2.

(c) Mean $\mu(s_2)$

(d) Slope $m(s_1)$.

Figure 4: Histograms with values for t_0, t_2, $m(s_1)$ and $\mu(s_2)$ features computed over the SEG supermarkets (Winter 2017 profiles).

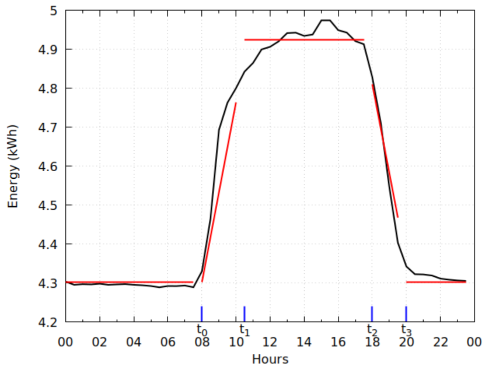
values (kWh) in Figure 4c and slopes values (kWh/h) in Figure 4d need to be used as they are continuous variables. Nine different intervals are created for the histograms and an additional bucket with the extreme values. Both average demand values for peak and off-peak periods show an important variability in their respective values. One reason for this large range of demand values is the large variability of the floor area. These two histograms are not normally distributed.

The same analysis can be computed for any season, year and store type. Results would be similar as the variability of the shape of the profiles is not very high. This is demonstrated when computing the errors. The $\overline{NP}$ evaluator between the real the reconstructed profiles (all years, seasons and set of stores) are computed. For the SE stores, the best $\overline{NP}$ results are 5.8%, 4.5% and 5.3% for 2017 Winter, Summer and Spring/Autumn ELDPs respectively. For the SEG stores, the best $\overline{NP}$ results are 4.3%, 4.1% and 4.1% for 2017 Winter, Summer and Spring/Autumn ELDPs respectively. We note that the error increases when the profiles are computed over older years. The worst $\overline{NP}$ scores is 7.2% computed over stores just with electricity over the Winter 2014 profiles. Comparing seasons, errors over Winter profiles are always slightly greater than for Spr/Aut profiles and errors over these ones are greater than for Summer profiles. The error for stores that consume electricity and gas is lower than stores than consume only electricity. This indicates that the heating system increases the complexity of the profile making the approximation of it using the proposed features more difficult. In analysing the shape of the profiles, we see that the demand fluctuations during the main peak are more common in Winter than Summer profiles, *e.g.* the 10am peak in Figure 2, or the afternoon in Figure 1. These fluctuations increase the error when modelling the consumption by averaging the demand over long periods, as we do with the reconstructed profile.

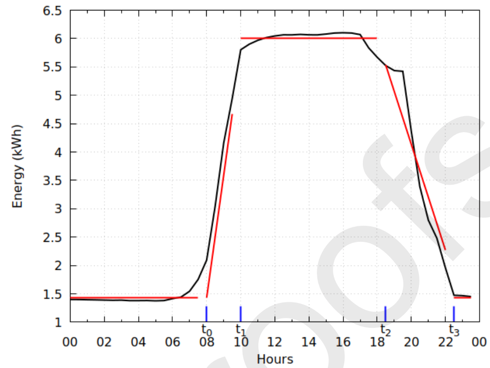
When representing the retail stores using the proposed feature and all the seasonal data $\overline{ED} = 1.0kWh$ and $\overline{NP}$=3.8%. Errors for this data-set are lower than errors obtained with the supermarkets as they have lower demand and a more regular U-inverted shape. Figure 5 displays the real and re-computed EDLPs for the case with the lowest NP (0.5%), median NP (3.5%) and worst NP (11.7%). The reconstructed EDLP in Figure 5a and Figure 5b match quite well the respective real EDLP. In the case of Figure 5c, the error is greater as there is an additional peak in the peak period and a valley in the off-peak period. Our model does not represent properly such events, but this type of event is unusual. Similar scores can be seen when using the Summer, Winter and Spring/Autumn profile.

4.2. Prediction experiments

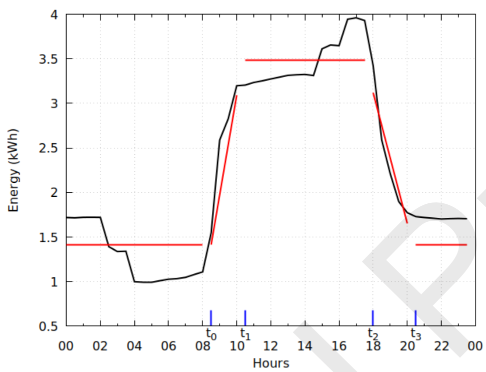
Prediction experiments are independently performed for all supermarket EDLPs computed during each year (2013-2017), season (Winter, Summer and Spring/Atumns) and store type (SE and SEG) giving a total of 5*3*2=30 different sets. An example of prediction for a particular supermarket (the example in Figure 2) is shown in Figure 6. The black profile is the real demand, the red and green profiles are predicted using the feature representation. These were the best predictions (considering the parameter search that minimises $\overline{ED}$ over all the set of stores) and they were obtained using



(a) Lowest NP.



(b) Median NP.



(c) Greatest NP

Figure 5: Real and reconstructed EDLP using the features with the lowest, median and worst NP scores for the retail store data-set.

OLS with features={GM area, Cafeteria area, Sales area, Office area, Chilled area} and k=98 for the whole profile representation and features={GM area, Cafeteria area, Sales area, Storage area, Chilled area, Location} and k=75 for the key feature representation. The values for the evaluators are ED=64.0 kWh and NP=16.6% when predicting the features and ED=59.5 kWh and 15.6% when predicting the whole profile. In this case, using the features implies a relative increase of the error of 7.5% and 6.4% with respect to the whole profile prediction for ED and NP evaluators, respectively.

Table 1 shows the results for the $\overline{NP}$ evaluator obtained when averaging the evaluator over all the supermarkets in the set.

The lowest error for the $\overline{NP}$ evaluator is 12.5% (Summer 2017) using

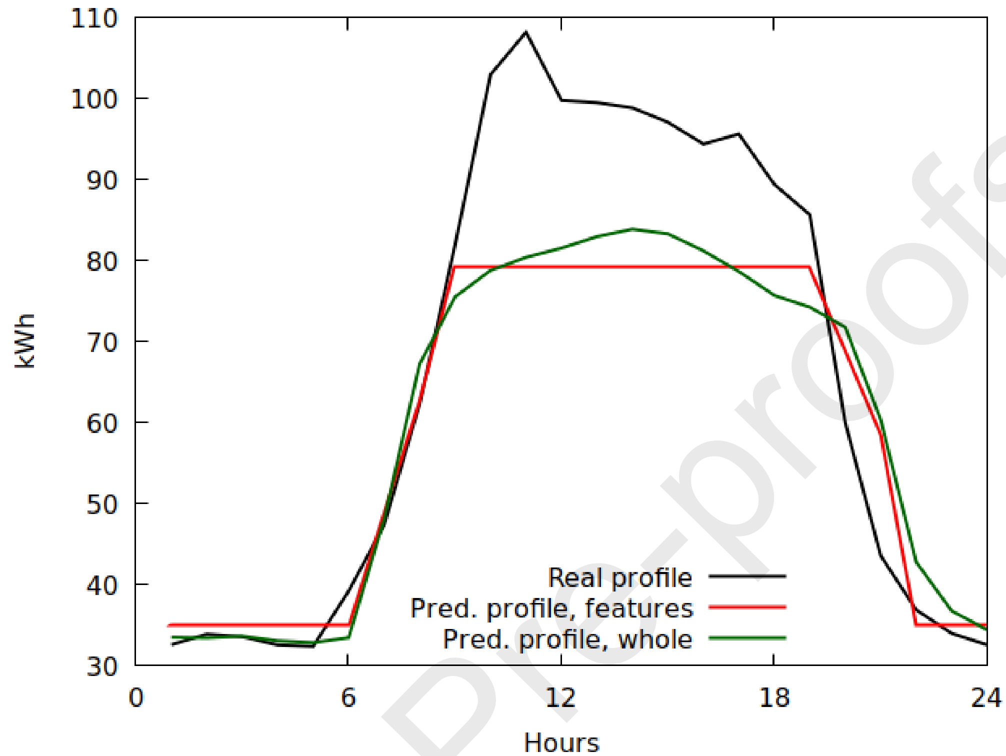


Figure 6: Real and predicted profiles using both the feature representation and the whole profile for one supermarket.

the OLS regression method for supermarkets using electricity and gas. This result is in line with those for the whole profile [13] and can be summarised thus:

- Errors computed over cold seasons are greater than errors obtained during warm seasons *i.e.* Summer profiles are better predicted than Spring/Autumn profiles, which are better than Winter profiles. The most likely cause is the uncertainty and variability of the heating system consumption:

- Errors obtained during most recent years are usually smaller than for old data. We suggest that stores tend to become more homogeneous as older appliances are routinely replaced.

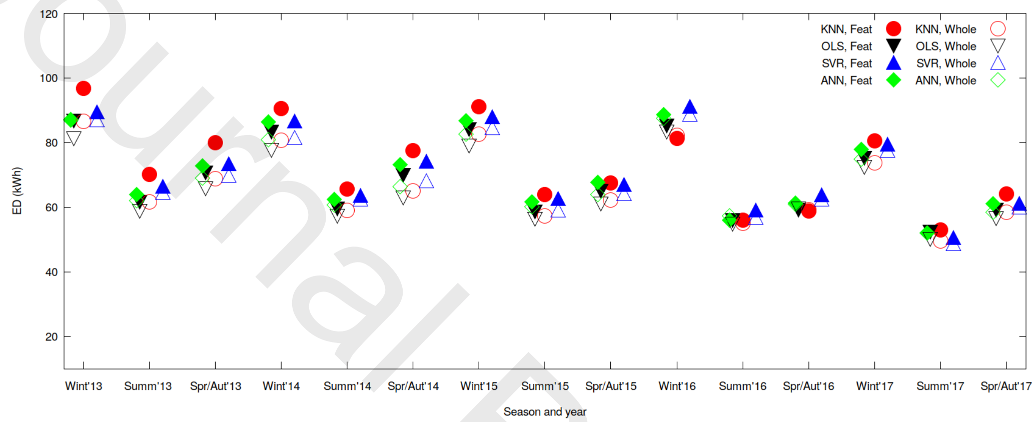
- There are only small differences when comparing algorithms. However, the OLS usually outperforms the other three regression methods which is due to the modest size of the data-sets.

- Stores with electricity and gas are better predicted than stores using electricity only. This too relates to the level of complexity added by the need to also predict the heating demand.

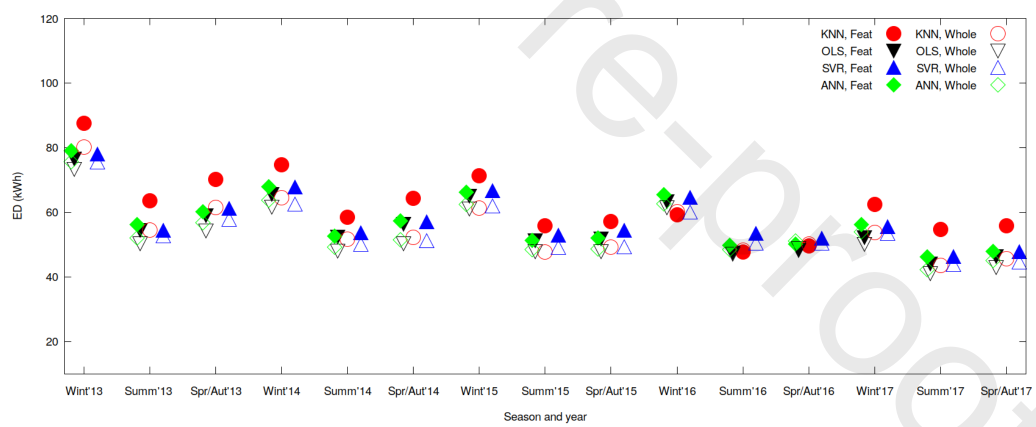
Comparing the results obtained using the feature set and those using whole profile representations shows the feasibility of exploiting reduced dimensionality to predict EDLPs. Figure 7 shows the $\overline{ED}$ values using both representations. The scores when using the full dimensional set (the whole profile) to predict the ELDP are better than using the reduced feature set. However, in many cases the difference is insignificant, especially for the most recent years. Using the $\overline{ED}$ evaluator the absolute difference is an average of 4.0 kWh (6.0%) and 4.4 kWh (8.3%) for SE and SEG, respectively, when comparing the two methods. For both SE and SEG, $\overline{NP}$ using the feature set is 0.9 points worse than using the whole profile. The relative differences for this evaluator are 4.6% and 5.9% for SE and SEG respectively.

TypSt	Year	Season	KNN	OLS	SVR	ANN
Stores just with elec. (SE)	2013	Wint	23.5	22.0	21.2	22.5
		Sum	20.8	18.9	19.4	19.6
		Spr/Aut	22.1	19.3	19.4	20.3
	2014	Wint	23.2	21.9	22.6	23.2
		Sum	20.6	19.2	20.2	20.5
		Spr/Aut	24.9	21.4	22.9	22.4
	2015	Wint	25.1	22.7	23.9	23.3
		Sum	23.0	20.2	20.9	21.5
		Spr/Aut	21.8	20.6	20.9	21.4
	2016	Wint	25.2	26.3	27.9	27.4
		Sum	19.7	18.6	18.8	19.6
		Spr/Aut	19.0	19.0	19.6	20.3
	2017	Wint	22.9	21.9	22.8	23.0
		Sum	17.7	18.1	17.6	19.2
		Spr/Aut	21.2	19.6	19.9	20.2
Stores with elec. and gas (SEG)	2013	Wint	21.5	18.5	18.9	19.3
		Sum	16.3	13.9	13.9	14.3
		Spr/Aut	17.9	15.2	15.8	15.5
	2014	Wint	19.9	17.1	17.9	18.6
		Sum	16.3	14.9	14.9	14.9
		Spr/Aut	17.3	15.6	15.9	15.8
	2015	Wint	18.7	17.4	17.9	17.9
		Sum	16.1	15.0	15.5	15.1
		Spr/Aut	16.2	14.7	15.6	15.3
	2016	Wint	17.2	17.7	18.1	18.6
		Sum	13.6	13.1	14.9	13.7
		Spr/Aut	14.3	13.5	14.4	14.1
	2017	Wint	17.5	14.6	15.6	16.2
		Sum	15.3	12.5	13.1	13.2
		Spr/Aut	16.0	13.1	13.7	13.9

Table 1: Prediction results using the $\overline{NP}$ (%) evaluator for the profile represented with the key features. Results are separated by algorithms, seasons, years and store type.



(a) Stores just with electricity (SE)



(b) Stores with electricity and gas (SEG)

Figure 7: Prediction results evaluated using the $\overline{ED}$ (kWh) for the profile represented with the reduced feature set (solid colour) and the whole profile. Points are offset in the x-axis for clarity.

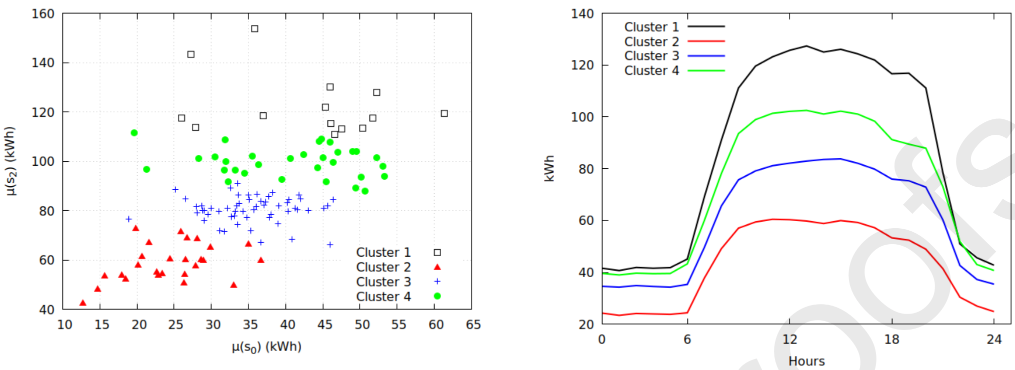
To understand the reasons for the greater error using reduced dimensionality it is necessary to re-think the sequence of processes performed in this prediction experiments (Figure 3). In this sequence, both modelling and prediction errors can occur throughout in the process chain. First, the profile to be predicted is modelled using the features with non-trivial error (see Section 4.1). Secondly, like any prediction process the features of the EDLP are not perfectly estimated using the regression model. Thirdly, when reconstructing the profile using these predicted features we are again approximating the whole profile adding new error.

As the evaluation is performed against the (full dimensional) real profile it seems logical to have greater error than predicting directly whole profiles. On the other hand, we have shown that the features are able to explain and capture the main patterns of the load profile with fewer parameters to predict than using the whole profile. Interestingly, as the difference in the results are small, the positive factors compensate the negative ones indicating the feasibility of using reduced dimensionality.

4.3. Clustering experiments

Clustering experiments are performed independently for all supermarket EDLPs computed during each year, season, and store type (SE, SEG and both together); a total of 5*3*3=45 experiments. Figure 8 shows the results obtained when clustering the EDLPs only represented with $\mu(s_0)$ and $\mu(s_2)$ (2-feat) when using readings during Winter 2017 of SEG supermarkets and the k-means algorithm (k=4). The clusters show a clear separation (Figure 8a), especially in the $\mu(s_2)$ feature because the value of $\mu(s_2)$ is greater than $\mu(s_0)$, giving more weight when computing distances among clusters. The real EDLPs of each cluster are used to compute the evaluators. The profile of each cluster centroid (Figure 8b) are distinct for both peak and off-peak periods.

To enable comparison, we computed the median with error bars using 95% confidence intervals using bootstrapping over all 45 experiments. Figure 9 shows these results over the supermarket data-set using the k-means for each one of the representations (whole profile, 8-feat, 4-feat and 2-feat). The results show only small differences between 2-feat clustering compared with using the whole profile. Interestingly, for the CDI (Figure 9a) and SI (Figure 9b) evaluators the clustering 2-feat results outperform those obtained with the whole profile when the number of clusters is greater than



(a) Clusters (b) Centroids using the EDLPs

Figure 8: Clustering results for EDLPs represented with $\mu(s_0)$ and $\mu(s_2)$ (only) using data for Winter 2017 of SEG supermarkets with k-means (k=4). Clusters 1, 2, 3 and 4 have 15, 26, 57 and 31 points, respectively.

three. Generally, 2-feat scores are better than scores obtained with 4-feat and 8-feat.

The fact that the 8-feat results include $\vec{t}$ are worse than the other clustering results is due to two factors: 1) $\vec{t}$ are numeric variables but they represent time intervals that are not well modelled by clustering algorithms that use Euclidean distances, and 2) the time intervals may add noise when creating the clusters as they are evaluated only using the demand differences of the whole profile.

Clustering results are given in Table 2 for all the evaluators averaged over all the whole profile (left value) and 2-feat (right value) experiments and number of cluster separated by algorithm. The differences between the values are small meaning that the results with both representations are similar. It might be expected that the whole-profile clustering evaluator would be better than the 2-feat results, however, for some algorithms and evaluators *e.g.* k-means and SI, or single link and SI, this is not the case.

For the retail store data-set, clustering experiments are performed independently for all the EDLPs computed during each season and for the whole year (Figure 5) with similar characteristics to the supermarkets data-set. When the number of clusters is small (less than four or five) the differences between the scores obtained with the whole profile and the reduced feature representation is greater than when using more clusters. Results obtained

Alg/Eval	CDI	MDI	SI	DBI	MIA	VRC
Kmeans	0.44/0.43	1.26/1.26	24.98/23.97	1.07/1.14	10.59/10.78	134.30/126.02
Single	0.30/0.30	1.14/1.25	6.85/6.64	0.55/0.57	8.88/9.19	10.92/14.47
Complete	0.35/0.36	0.76/0.88	14.89/15.79	0.93/1.00	10.06/10.41	116.63/107.78
UPGMA	0.30/0.31	0.57/0.70	9.29/9.40	0.72/0.83	9.34/9.82	90.40/91.02
WPGMA	0.32/0.33	0.65/0.77	13.57/12.78	0.82/0.90	9.50/9.97	96.47/94.18
UPGMC	0.28/0.29	0.56/0.69	8.99/8.83	0.65/0.78	8.94/9.64	84.64/87.24
WPGMC	0.28/0.30	0.60/0.69	9.82/10.09	0.65/0.80	8.98/9.61	80.44/90.69
WARD	0.47/0.49	1.36/1.51	29.01/29.37	1.14/1.23	10.59/10.98	126.54/118.02

Table 2: Clustering results for the supermarket data-set for all evaluators averaged over all the whole profile (left value), 2-feat (right value), and number of cluster separated by the algorithm.

with 8-feat are consistently worse than those obtained with the other representations. Clustering results are given in Table 3 or all the evaluators averaged over all the whole profile (left value) and 2-feat (right value) experiments and number of cluster separated by algorithm. The results obtained with the whole profile marginally outperform those obtained with the 2-feat, with exceptions such a UPGM algorithm and SI evaluator.

Alg/Eval	CDI	MDI	SI	DBI	MIA	VRC
Kmeans	0.21/0.24	2.90/2.56	13.72/17.06	0.87/0.86	1.66/1.77	750.69/734.90
Single	0.09/0.09	0.52/0.66	2.93/2.91	0.21/0.27	0.85/0.91	141.50/145.86
Complete	0.14/0.17	0.72/0.83	3.96/4.71	0.59/0.66	1.35/1.44	497.56/507.80
UPGMA	0.12/0.12	0.44/0.51	3.34/3.38	0.44/0.49	1.19/1.20	278.29/342.59
WPGMA	0.12/0.13	0.54/0.70	3.46/3.49	0.50/0.53	1.19/1.24	370.92/381.44
UPGMC	0.12/0.12	0.45/0.49	3.41/3.13	0.44/0.47	1.16/1.17	308.14/333.41
WPGMC	0.13/0.13	0.51/0.53	3.62/3.64	0.46/0.49	1.24/1.23	270.93/352.74
WARD	0.66/0.79	5.09/7.81	143.51/139.43	1.21/1.31	2.05/2.47	484.12/388.30

Table 3: Clustering results for the retail stores data-set for all evaluators averaged over all the whole profile (left value) and 2-feat (right value) experiments, and number of cluster separated by algorithm.

As a final remark about the clustering results, evaluation scores for the 2-feat clustering results are slightly worse than those obtained when using the whole profile when using less than four clusters. However, evaluation scores for these two representations are very close when the number of clusters is greater than four or averaged over the total number of clusters. The 2-feat works well for clustering the profiles because these two features ($\mu(s_0)$ and $\mu(s_2)$) are the main behavioural drivers accounting for most of the EDLP.

5. Conclusions and future work

Our aim was to investigate whether dimensional reduction could generate a statistically reasonable representation the EDLP of a retail store such that it could be used to predict the electricity demand for a new store in the portfolio of a company. Previously we have shown how this can be done using the whole profile, but a simpler representation of the values of the features (*e.g.* Figure 4) may offer advantages by reducing the complexity of the problem. In particular, whether it could help detect trends and anomalous behaviours within EDLPs.

We have studied the impact to reduced-feature sets to represent EDLPs for prediction and clustering using real data of two distinct data-sets: supermarkets with 1-h resolution readings (prediction and clustering) and retail stores with 30-min resolution (clustering only). We have demonstrated that the extracted features give a good description of the original EDLP *i.e.* being able to re-construct the EDLP with only a small error. However, we need to be aware that for a small number of stores (*e.g.* Figure 5c) the proposed representation did not work so well. In general though, we have shown that the evaluation scores are the same or only marginally worse than results obtained using the whole profile. The results are robust as the two tasks are different in nature: prediction is supervised learning meanwhile clustering is unsupervised.

This proposed simplified representation is a more concise way to represent the EDLP than using the whole EDLP (real resolution values). For some types of analyses, small variances of the demand within the time period can be considered superfluous information that does not add useful information to the overall picture. For example, as the repeated night-time demand values (Figure 1) are repeated over a long period, using an average value is sufficient to summarise and represent the demand during these periods.

The main implication for energy managers and researchers is that a reduced number of features is easier to interpret and visualise instead of a high resolution EDLP. The clustering results suggest its utility as dimensional reduction technique to cope with the ‘curse of dimensionality’. More generally, we have demonstrated that a simpler way to represent data can work as well for some specific energy problems as complex and high resolution representation. As modern (networked) sensors increase the volume, availability, and immediacy, transforming such high-resolution data streams in a ‘smart‘ way based on observed behaviours may be helpful. The proposed features to represent the EDLP may have limitations for applications such as investigating and predicting demand shifting and demand variability for energy management purposes . This is due to the lack of granularity which will not allow detection of demand changes at specific times (e.g. hourly).

As future work, we suggest that reduced-feature representation can be applied to any electricity data-set of retail facilities with a diurnal opening schedule. Moreover, this feature-reduction technique can be applied classification. Furthermore, combining both clustering and prediction may be an interesting approach to separately predict the demand by existing buildings that are in each cluster. This is different to predicting the demand of new buildings, but large data-sets in both temporal dimensional and number of stores would be required for such analysis.

References

[1] UNFCCC, COP26 The Glasgow Climate Pact, Glasgow, UK (2021). URL https://ukcop26.org/wp-content/uploads/2021/11/COP26-Presidency-Outcomes-The-Climate-Pact.pdf

[2] European Comission, A Clean Planet for all: A European strategic long-term vision for a prosperous, modern, competitive and climate neutral economy, (No. COM(2018) 773 final). Brussels, Belgium (2018). URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0773&from=EN

[3] CCC, Reducing UK emissions: 2019 Progress Report to Parliament. Committee on Climate Change, London, UK (2019). URL https://www.theccc.org.uk/wp-content/uploads/2019/07/CCC-2019-Progress-in-reducing-UK-emissions.pdf

[4] J. Rogelj, D. Shindell, K. Jiang, S. Fifita, P. Forster, V. Ginzburg, C. Handa, H. Kheshgi, S. Kobayashi, E. Kriegler, L. Mundaca, R. Se-

ferian, M. Vilarino, K. Calvin, O. Edelenbosch, J. Emmerling, S. Fuss, T. Gasser, N. Gillet, C. He, E. Hertwich, L. H. Isaksson, D. Huppmann, G. Luderer, A. Markandya, D. McCollum, R. Millar, M. Meinshausen, A. Popp, J. Pereira, P. Purohit, K. Riahi, A. Ribes, H. Saunders, C. Schadel, C. Smith, P. Smith, E. Trutnevyte, Y. Xiu, K. Zickfeld, W. Zhou, Chapter 2: Mitigation pathways compatible with 1.5 °C in the context of sustainable development, in: Global Warming of 1.5 °C an IPCC special report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, Intergovernmental Panel on Climate Change, 2018, pp. 93–174.
URL http://pure.iiasa.ac.at/id/eprint/15515/

[5] E. Bassas, J. Patterson, P. Jones, A review of the evolution of green residential architecture, Renewable and Sustainable Energy Reviews 125 (2020) 109796.
URL https://doi.org/10.1016/j.rser.2020.109796

[6] H. Rau, P. Moran, R. Manton, J. Goggins, Changing energy cultures? Household energy use before and after a building energy efficiency retrofit, Sustainable Cities and Society 54 (2020) 101983.
URL https://doi.org/10.1016/j.scs.2019.101983

[7] E. Cuce, An overview of domestic energy consumption in the UK: past, present and future, International Journal of Ambient Energy (2014) 1–8.
URL https://doi.org/10.1080/01430750.2014.973120

[8] R. Granell, C. Axon, D. Wallom, R. Layberry, Power-use profile analysis of non-domestic consumers for electricity tariff switching, Energy Efficiency 9 (2016) 825–841.
URL https://doi.org/10.1007/s12053-015-9404-9

[9] C. Axon, S. Bright, T. Dixon, K. Janda, M. Kolokotroni, Building communities: Reducing energy use in tenanted commercial property, Building Research and Information 40 (2012) 461–472.
URL https://doi.org/10.1080/09613218.2012.680701

[10] F. M. Dahunsi, A. E. Olawumi, D. T. Ale, O. A. Sarumi, A systematic review of data pre-processing methods and unsupervised mining meth-

ods used in profiling smart meter data, AIMS Electronics and Electrical Engineering 5 (4) (2021) 284–314.
URL https://doi.org/10.3934/electreng.2021015

[11] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, Energy Vol. 42 (1) (2012) 68–80.
URL https://doi.org/10.1016/j.energy.2011.12.031

[12] S. Yilmaz, J. Chambers, M. Patel, Comparison of clustering approaches for domestic electricity load profile characterisation - implications for demand side management, Energy 180 (2019) 665–677.
URL https://doi.org/10.1016/j.energy.2019.05.124

[13] R. Granell, C. Axon, M. Kolokotroni, D. Wallom, Predicting electricity demand profiles of new supermarkets using machine learning, Energy and Buildings 234 (2021) 110635–110635.
URL https://doi.org/10.1016/j.enbuild.2020.110635

[14] C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, 2006.

[15] R. E. Bellman, Adaptive control processes, Princeton University Press, 2015.

[16] W. Chung, Review of building energy-use performance benchmarking methodologies, Applied Energy 88 (5) (2011) 1470 – 1479.
URL http://doi.org/10.1016/j.apenergy.2010.11.022

[17] Z. Li, Y. Han, P. Xu, Methods for benchmarking building energy consumption against its past or intended performance: An overview, Applied Energy 124 (2014) 325 – 334.
URL http://doi.org/10.1016/j.apenergy.2014.03.020

[18] H.-X. Zhao, F. Magouls, A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews 16 (6) (2012) 3586 – 3592.
URL https://doi.org/10.1016/j.rser.2012.02.049

[19] C. Deb, F. Zhang, J. Yang, S. E. Lee, K. W. Shah, A review on time series forecasting techniques for building energy consumption, Renewable

and Sustainable Energy Reviews 74 (2017) 902 – 924.
URL https://doi.org/10.1016/j.rser.2017.02.085

[20] M. Bourdeau, X.-Q. Zhai, E. Nefzaoui, X. Guo, P. Chatellier, Modeling and forecasting building energy consumption: A review of data-driven techniques, Sustainable Cities and Society 48 (2019) 101533.
URL https://doi.org/10.1016/j.scs.2019.101533

[21] K. Li, W. Xue, G. Tan, A. S. Denzer, A state of the art review on the prediction of building energy consumption using data-driven technique and evolutionary algorithms, Building Services Engineering Research and Technology 41 (1) (2020) 108–127.
URL https://doi.org/10.1177/0143624419843647

[22] K. Yun, R. Luck, P. J. Mago, H. Cho, Building hourly thermal load prediction using an indexed arx model, Energy and Buildings 54 (2012) 225–233.
URL https://doi.org/10.1016/j.enbuild.2012.08.007

[23] K. Jeong, C. Koo, T. Hong, An estimation model for determining the annual energy cost budget in educational facilities using sarima (seasonal autoregressive integrated moving average) and ann (artificial neural network), Energy 71 (2014) 71–79.
URL https://doi.org/10.1016/j.energy.2014.04.027

[24] D. W. Schrock, D. E. Clarige, Predicting energy usage in a supermarket, in: Proceedings of the Sixth Symposium on Improving Building Systems in Hot and Humid Climates, 1989, pp. 19–27.

[25] K. Lindberg, S. Bakker, I. Sartori, Modelling electric and heat load profiles of non-residential buildings for use in long-term aggregate load forecasts, Utilities Policy 58 (2019) 63–88.
URL https://doi.org/10.1016/j.jup.2019.03.004

[26] M. S. Spyrou, K. Shanks, M. J. Cook, J. Pitcher, R. Lee, An empirical study of electricity and gas demand drivers in large food retail buildings of a national organisation, Energy and Buildings 68, Part A (2014) 172 – 182.
URL http://doi.org/10.1016/j.enbuild.2013.09.015

[27] W. Chung, Y. Hui, Y. M. Lam, Benchmarking the energy efficiency of commercial buildings, Applied Energy 83 (1) (2006) 1 – 14.
URL http://doi.org/10.1016/j.apenergy.2004.11.003

[28] M. Braun, H. Altan, S. Beck, Using regression analysis to predict the future energy consumption of a supermarket in the UK, Applied Energy 130 (2014) 305 – 313.
URL http://doi.org/10.1016/j.apenergy.2014.05.062

[29] O. Valgaev, F. Kupzog, Building power demand forecasting using k-nearest neighbors model - initial approach, in: 2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), 2016, pp. 1055–1060.

[30] Z. Ma, J. Song, J. Zhang, Energy consumption prediction of air-conditioning systems in buildings by selecting similar days based on combined weights, Energy and Buildings 151 (2017) 157 – 166.
URL https://doi.org/10.1016/j.enbuild.2017.06.053

[31] M. W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy and Buildings 147 (2017) 77 – 89.
URL https://doi.org/10.1016/j.enbuild.2017.04.038

[32] G. K. Tso, K. K. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy 32 (9) (2007) 1761–1768.
URL https://doi.org/10.1016/j.energy.2006.11.010

[33] Y. T. Chae, R. Horesh, Y. Hwang, Y. M. Lee, Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings, Energy and Buildings 111 (2016) 184–194.
URL https://doi.org/10.1016/j.enbuild.2015.11.045

[34] C. Deb, L. S. Eang, J. Yang, M. Santamouris, Forecasting diurnal cooling energy load for institutional buildings using artificial neural networks, Energy and Buildings 121 (2016) 284–297.
URL https://doi.org/10.1016/j.enbuild.2015.12.050

[35] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Predicting hourly cooling load in the building: A comparison of support vector machine and

different artificial neural networks, Energy Conversion and Management 50 (1) (2009) 90 – 96.
URL https://doi.org/10.1016/j.enconman.2008.08.033

[36] D. Zhao, M. Zhong, X. Zhang, X. Su, Energy consumption predicting model of vrv (variable refrigerant volume) system in office buildings based on data mining, Energy 102 (2016) 660 – 668.
URL https://doi.org/10.1016/j.energy.2016.02.134

[37] D. Datta, S. Tassou, D. Marriott, Application of neural networks for the prediction of the energy consumption in a supermarket, Proceedings of the Clima 2000 Conference, Brussels, Belgium.

[38] G. Revati, J. Hozefa, S. Shadab, A. Sheikh, S. R. Wagh, N. M. Singh, Smart building energy management: Load profile prediction using machine learning, in: 2021 29th Mediterranean Conference on Control and Automation (MED), 2021, pp. 380–385. doi:10.1109/MED51440.2021.9480170.

[39] G. Hafeez, K. S. Alimgeer, I. Khan, Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid, Applied Energy 269 (2020) 114915.
URL https://doi.org/10.1016/j.apenergy.2020.114915

[40] X. Luo, L. O. Oyedele, A. O. Ajayi, O. O. Akinade, H. A. Owolabi, A. Ahmed, Feature extraction and genetic algorithm enhanced adaptive deep neural network for energy consumption prediction in buildings, Renewable and Sustainable Energy Reviews 131 (2020) 109980.
URL https://doi.org/10.1016/j.rser.2020.109980

[41] B. Dong, C. Cao, S. E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings 37 (5) (2005) 545 – 553.
URL https://doi.org/10.1016/j.enbuild.2004.09.009

[42] S. Paudel, M. Elmitri, S. Couturier, P. H. Nguyen, R. Kamphuis, B. Lacarrire, O. Le Corre, A relevant data selection method for energy consumption prediction of low energy building based on support vector machine, Energy and Buildings 138 (2017) 240 – 256.
URL https://doi.org/10.1016/j.enbuild.2016.11.009

[43] R. E. Edwards, J. New, L. E. Parker, Predicting future hourly residential electrical consumption: A machine learning case study, Energy and Buildings 49 (2012) 591–603.
URL https://doi.org/10.1016/j.enbuild.2012.03.010

[44] R. K. Jain, K. M. Smith, P. J. Culligan, J. E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Applied Energy 123 (2014) 168–178.
URL https://doi.org/10.1016/j.apenergy.2014.02.057

[45] G. J. Tsekouras, N. D. Hatziargyriou, E. N. Dialynas, Two-stage pattern recognition of load curves for classification of electricity customers, IEEE Transactions on Power Systems 22 (3) (2007) 1120–1128.
URL https://doi.org/10.1109/TPWRS.2007.901287

[46] P. Nystrup, H. Madsen, E. M. Blomgren, G. de Zotti, Clustering commercial and industrial load patterns for long-term energy planning, Smart Energy 2 (2021) 100010.
URL https://doi.org/10.1016/j.segy.2021.100010

[47] G. Chicco, R. Napoli, F. Piglione, Comparisons among clustering techniques for electricity customer classification, IEEE Transactions on Power Systems 21 (2) (2006) 933–940.
URL https://doi.org/10.1109/TPWRS.2006.873122

[48] R. Granell, C. J. Axon, D. C. Wallom, Clustering disaggregated load profiles using a Dirichlet process mixture model, Energy Conversion and Management 92 (2015) 507–516.
URL http://dx.doi.org/10.1016/j.enconman.2014.12.080

[49] A. Notaristefano, G. Chicco, F. Piglione, Data size reduction with symbolic aggregate approximation for electrical load pattern grouping, IET Generation, Transmission Distribution 7 (2) (2013) 108–117.

[50] R. Granell, C. J. Axon, D. C. H. Wallom, Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles, IEEE Transactions on Power Systems 30 (6) (2015) 3217–3224.
URL https://doi.org/10.1109/TPWRS.2014.2377213

[51] S. Roberts, J. Thunim, Managing and mining smart meter data - at scale, cSE Project Showcase Presentation. Centre for Sustainable Energy (2013).

[52] Z. Mylona, M. Kolokotroni, S. A. Tassou, Frozen food retail: Measuring and modelling energy use and space environmental systems in an operational supermarket, Energy and Buildings 144 (2017) 129 – 143.
URL `http://doi.org/10.1016/j.enbuild.2017.03.049`

[53] C. Cortes, V. Vapnik, Support-vector networks, in: Machine Learning, 1995, pp. 273–297.

[54] F. Hayashi, Econometrics, Princeton University Press, 2011.

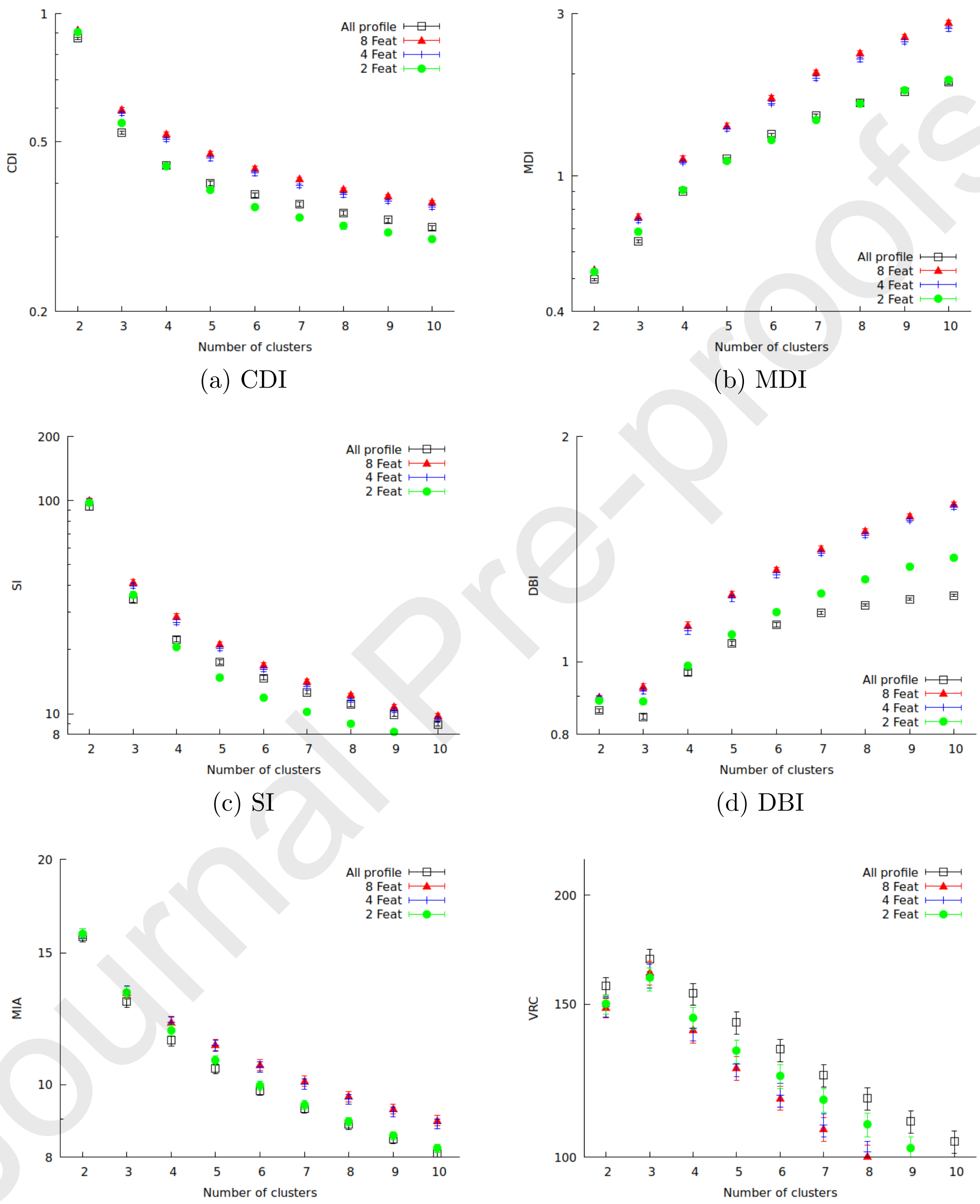
[55] S. Fritsch, F. Guenther, Neuralnet: Training of Neural Networks, R package version 1.33 (2016).
URL `https://CRAN.R-project.org/package=neuralnet`

[56] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, r package version 1.6-8 (2017).
URL `https://CRAN.R-project.org/package=e1071`

[57] J. Williams, Clustering household electricity use profiles, in: Proceedings of MLSDA, ACM, 2013, pp. 19–26.

[58] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, M. Kolehmainen, Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data, Applied Energy 87 (11) (2010) 3538–3545.
URL `https://doi.org/10.1016/j.apenergy.2010.05.015`

[59] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2-3) (2001) 107–145.



(a) CDI (b) MDI

(c) SI (d) DBI

(e) MIA (f) VRC

Figure 9: Clustering results for the supermarket data-set using the k-means. N.B. the Y-axis is log scale.

Declaration of interests

☒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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: