Article

Smart Meters Time Series Clustering for Demand Response Applications in the Context of High Penetration of Renewable Energy Resources

Santiago Bañales [1,2], Raquel Dormido [1,*] and Natividad Duro [1]

[1] Department of Computer Sciences and Automatic Control, Universidad Nacional de Educación a Distancia (UNED), C/Juan del Rosal, 16, 28015 Madrid, Spain; sbanales2@alumno.uned.es (S.B.); nduro@dia.uned.es (N.D.)
[2] Iberdrola Innovation Middle East, Doha 210177, Qatar
* Correspondence: raquel@dia.uned.es

Abstract: The variability in generation introduced in the electrical system by an increasing share of renewable technologies must be addressed by balancing mechanisms, demand response being a prominent one. In parallel, the massive introduction of smart meters allows for the use of high frequency energy use time series data to segment electricity customers according to their demand response potential. This paper proposes a smart meter time series clustering methodology based on a two-stage k-medoids clustering of normalized load-shape time series organized around the day divided into 48 time points. Time complexity is drastically reduced by first applying the k-medoids on each customer separately, and second on the total set of customer representatives. Further time complexity reduction is achieved using time series representation with low computational needs. Customer segmentation is undertaken with only four easy-to-interpret features: average energy use, energy–temperature correlation, entropy of the load-shape representative vector, and distance to wind generation patterns. This last feature is computed using the dynamic time warping distance between load and expected wind generation shape representative medoids. The two-stage clustering proves to be computationally effective, scalable and performant according to both internal validity metrics, based on average silhouette, and external validation, based on the ground truth embedded in customer surveys.

Keywords: time series clustering; time series representation; electrical smart meters; demand response; renewable energy; clustering validation

Citation: Bañales, S.; Dormido, R.; Duro, N. Smart Meters Time Series Clustering for Demand Response Applications in the Context of High Penetration of Renewable Energy Resources. *Energies* **2021**, *14*, 3458. https://doi.org/10.3390/en14123458

Academic Editor: Djaffar Ould-Abdeslam

Received: 12 May 2021
Accepted: 9 June 2021
Published: 11 June 2021

1. Introduction

Demand response (DR), or the capability of electrical loads to adapt their shape at specific points in time given the right incentives, is receiving increasing attention from policy makers and energy system designers [1]. The key trends driving the development of the energy system, decarbonization, electrification and digitalization converge towards an increasing need for demand flexibility enhanced by digital technologies [2]. Notwithstanding the large uncertainties introduced by the COVID-19 pandemic in the energy system [3], a sustainable economic recovery is set to be based on channeling new investments in clean energy and further digitalization [4], which would further promote the automation of demand response programs [5], in particular for residential customers [6] in the context of higher penetration of renewable resources.

The global deployment of smart meters has also resulted in a dramatic increase in Artificial Intelligence (AI) and Machine Learning (ML) applications to improve the planning and operation of the power system under the big data paradigm [7–9], as customer segmentation for DR applications is one the most common applications [10]. Time series clustering approaches have been largely applied to smart meters load profiles datasets [11–13]. The challenges associated with time series clustering are well recognized, and they include high dimensionality and the definition of similarity taking the time dimension into account, from which three key research areas are derived: dimensionality reduction; clustering approach, which includes the choice of distance measurement, clustering prototypes and clustering algorithm; and clustering performance evaluation [14,15].

The most common approach for dimensionality reduction is to transform the smart meters times series data into a set of customer features that capture, according to a heuristic and expert-based criteria, the structure of the customer's load shape. These features may be related to DR flexibility metrics, such as entropy analysis and thermal profiling [16], an average energy aggregation over certain periods of time [17], a combination of mean load level at chosen time intervals and outdoor temperature ranges [18,19], daily energy, minimum and maximum active powers [20], or peak characterization (e.g., peak time, duration and intensity) [21]. Other approaches for dimensionality reduction include principal component analysis (PCA) [17,22], statistical parameters of the energy use probability distributions such as skewness and kurtosis [23], time series analysis such as autocorrelation [24], or deep-learning-based convolutional autoencoder (CAE) to reduce to a representative vector in the encoded space [25]. In [26], the authors undertake a systematic comparison between different dimensionality reduction techniques such as features-based (seasonal averages and maximums, seasonal median, maximum and median variation), non-data adaptive (piecewise aggregate approximation, discrete wavelet transform), data adaptive (piecewise linear approximation (PLA)), and model based (multiple linear regression, robust linear regression, generalized additive model, Holt–Winters exponential smoothing), finding that best results were achieved by model-based representations and the adaptive method PLA. K-centered clustering algorithms are the most popular, with k-means the most common prototype [12,13]. Hierarchical clustering (HC) is also fairly used, but normally applied to smaller datasets, and self-organized maps (SOM) is commonly used for its strong visualization features. Recent trends aim at using alternative clustering techniques such as density-based methods [27,28] and modeling the embedded uncertainty and indetermination of energy use data [29,30]. Multi-stage clustering has been used to increase performance by dealing in turn with absolute load and normalized load shape [31], and to deal with dimensionality by first computing local representatives at the customer level and then global representatives at the global level using a load shape dictionary (LSD) approach [32]. The dictionary approach, combined with an adaptive k-means clustering algorithm, has also been proven to be effective and scalable to large datasets [21,33]. Euclidean distance is the most used dissimilarity measure, while fast algorithms for dynamic time warping (DTW) distance are recommended when comparing raw smart meters time series data [13,32]. For clustering performance evaluation average silhouette and the Dunn index are commonly used as internal clustering validation metrics, and external validation is rarely used [12].

The goals of this research are two-fold. The first research objective is to develop a time series clustering methodology that takes explicitly into account renewable energy generation patterns. Although previous literature often describes the high impact of a high penetration of renewables in the development of DR mechanisms, this is the first time that this impact is quantitively embedded in the clustering methodology, to the knowledge of the authors. The second objective is to design a time series clustering strategy that is scalable and computationally efficient using a combination of techniques, such as multi-step clustering and dimensionality reduction.

The rest of the paper is structured as follows. Section 2 describes the methodology. Section 3 applies the methodology using the well-known public dataset from Irish CER smart meter trial. Section 4 discusses the results on the light of previous research. Section 5 summarizes the conclusions and points to further research.

2. Methodology

Figure 1 illustrates the methodology. The 5 steps in the top of the figure can be grouped into four blocks: data analysis, 2-step clustering, distance to wind and DR applications.

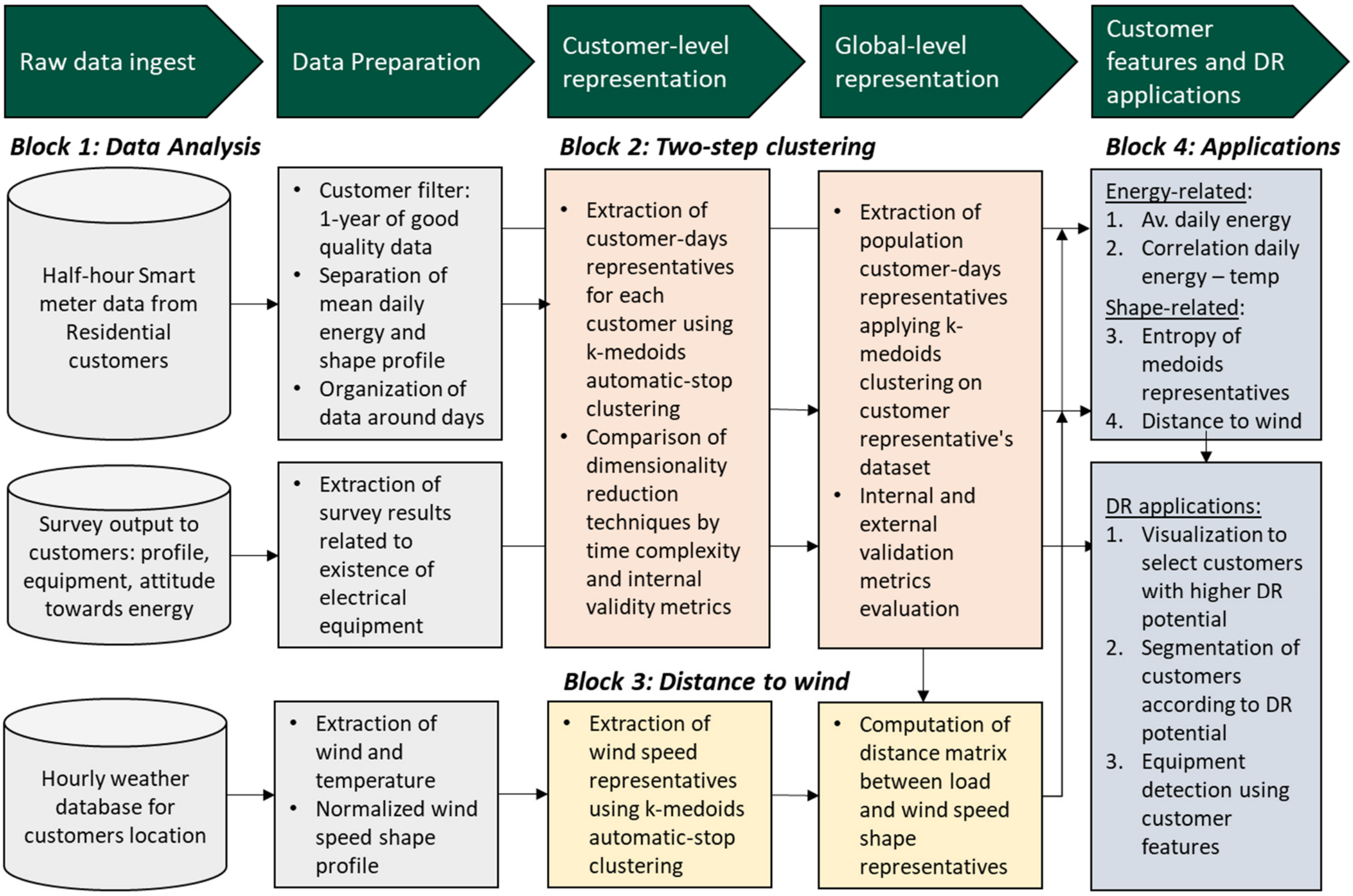
Raw data ingest
Data Preparation
Customer-level representation
Global-level representation
Customer features and DR applications
Block 1: Data Analysis
Block 2: Two-step clustering
Block 4: Applications
Half-hour Smart meter data from Residential customers
• Customer filter: 1-year of good quality data
• Separation of mean daily energy and shape profile
• Organization of data around days
• Extraction of customer-days representatives for each customer using k-medoids automatic-stop clustering
• Comparison of dimensionality reduction techniques by time complexity and internal validity metrics
• Extraction of population customer-days representatives applying k-medoids clustering on customer representative's dataset
• Internal and external validation metrics evaluation
Energy-related:
1. Av. daily energy
2. Correlation daily energy – temp
Shape-related:
3. Entropy of medoids representatives
4. Distance to wind
Survey output to customers: profile, equipment, attitude towards energy
• Extraction of survey results related to existence of electrical equipment
DR applications:
1. Visualization to select customers with higher DR potential
2. Segmentation of customers according to DR potential
3. Equipment detection using customer features
Block 3: Distance to wind
Hourly weather database for customers location
• Extraction of wind and temperature
• Normalized wind speed shape profile
• Extraction of wind speed representatives using k-medoids automatic-stop clustering
• Computation of distance matrix between load and wind speed shape representatives

Figure 1. Smart meters time series clustering for DR applications methodology.

2.1. Block 1: Data Analysis

Raw data are composed of three different but interconnected datasets: smart meters time series data (with hourly or half-hourly frequency) for each customer, a customer ground-truth set containing customer features related to customer profile, premises type and available equipment, and a weather information time series dataset (with hourly frequency) from the weather stations closest to each customer. The dataset is filtered to retain customers with a full year's worth of hourly smart meter data and a good level of quality (i.e., low number of missing values or clear outliers) and missing data are interpolated using standard techniques.

There are two key transformations of the data in this methodology: the breakdown of the smart meter dataset in daily energy use (in kWh/day) and normalized load shape (ratio of hourly energy and daily mean hourly energy) and the organization of the data for each customer around days. Hence, each customer is characterized by one vector of daily mean energy use of size equal to the number of days with data and a matrix of normalized load shape days, with as many rows as days and as many columns as number of daily data points (normally 24 or 48). Weather data time series are treated in a similar way, filtering the hourly time series data for wind speed and temperature for the selected one-year period, and then extracting two vectors of size equal to the number of days in the year, one of mean daily temperature and other the normalized wind speed, defined as the ratio of hourly wind speed and the daily average wind speed.

2.2. Block 2: Two-Step Clustering

An effective times series clustering for large datasets heavily depends on the application of successful reduction techniques and classification processes aiming at a reduction in high data dimensionality. Once data quality is ensured as output of block 1, the goal of block 2 is to drastically reduce the computational complexity of the problem. In this paper a very large dataset (e.g., for 5000 customers and one year worth of half-hourly smart meter data, the size of the distance matrix would be of 5000 customers × 365 days = ~1.8 million rows and 48 columns) is transformed to a lower dimensionality by first applying three different time series representation techniques, and second by using a two-step clustering procedure.

Three time series representation techniques have been chosen because of their relatively low computation complexity covering three different categories: piecewise aggregate approximation (PAA) (non-data adaptive category) [34], symbolic aggregate approximation (SAX) (data adaptive) [35] and features based on clipping (data dictated) [36]. *TSrep* package in language R [37] has been used to implement these techniques. To compute the dissimilarity matrix, Euclidean distance is used on normalized values for PAA and clipping and Gower's distance for the non-numeric symbolic approach SAX [38]. Second, a two-step k-medoid partitioning clustering approach is applied to the transformed dataset. K-medoid partitioning clustering has been widely used with success for smart meter datasets, and the use of medoids as protypes provides more robust results in terms of noise and outliers than average- or median-based prototypes, guarantees convergence, and allows for the use of different distance functions other than Euclidean [13]. The first step k-medoids clustering algorithm with automatic stop criteria is applied to the customer-day elements of each customer (365 elements for one year worth of data). The stop criterion is based on the computation of the average silhouette for each iteration of increasing number of clusters: the algorithm stops when the average silhouette decreases with the number of clusters. The medoids of each customer-day clustering are the representative of each one of the customers. The second k-medoids algorithm is applied to the set of population representatives medoids, resulting in the final selection of a few medoids to represent the whole population. Figure 2 summarizes the process of the data transformation and the multi-step clustering in flow-chart format. A mathematical notation of this process is also provided in Appendix A.

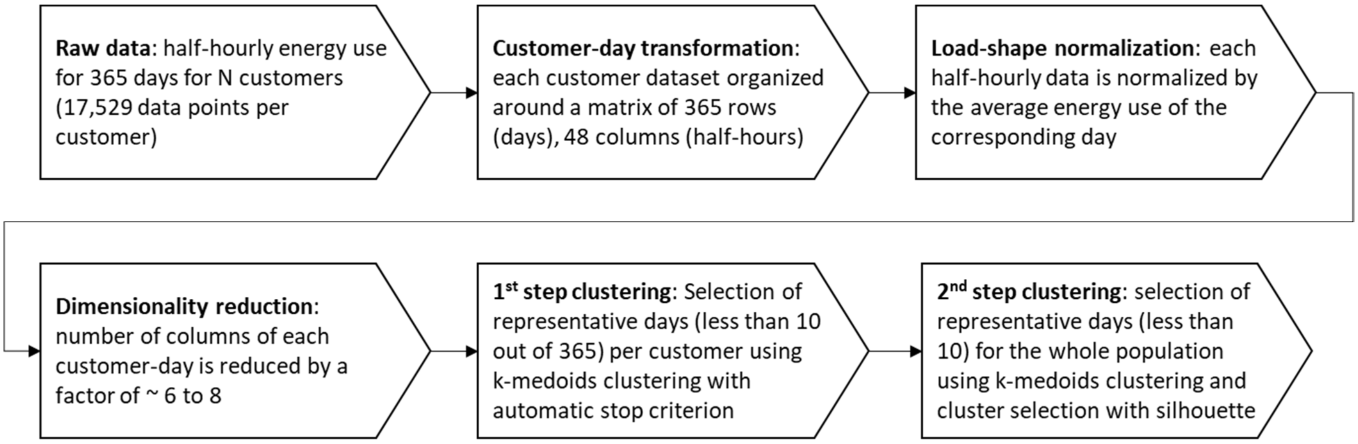
Raw data: half-hourly energy use for 365 days for N customers (17,529 data points per customer)
Customer-day transformation: each customer dataset organized around a matrix of 365 rows (days), 48 columns (half-hours)
Load-shape normalization: each half-hourly data is normalized by the average energy use of the corresponding day
Dimensionality reduction: number of columns of each customer-day is reduced by a factor of ~ 6 to 8
1st step clustering: Selection of representative days (less than 10 out of 365) per customer using k-medoids clustering with automatic stop criterion
2nd step clustering: selection of representative days (less than 10) for the whole population using k-medoids clustering and cluster selection with silhouette

Figure 2. Flow-chart of the data transformation and multi-step time series clustering process.

The different dimensionality reduction techniques are compared according to four metrics: computation time (for distance matrix and clustering), number of clusters generated by the automatic stop criterion algorithm, average silhouette, and percentage of negative silhouette values. Computation time provides a measure of the efficiency of the technique, the number of clusters is a measure of the capability of each technique to find subtle patterns in the shape of the data and the silhouette metrics (average and number of negative silhouette elements) provide a robust internal validation of the separation and compactness of the clustering results. Additionally, an external validation metric is also computed for the three methods based on the relevance of regression logistic coefficients predicting the probability of the existence of a given electrical equipment using the relative frequency of the second phase of clustering as predictors.

Time complexity of k-medoids algorithm is high compared to k-means since it has a quadratic dependence to the number of elements to cluster, or O(n(n-k)^2), where n is the number of elements and k the number of clusters [39]. The methodology proposed radically reduces the time complexity first via dimensionality reduction by a factor of $\left(\frac{h}{f}\right)^2$, where h is the number of periods per day (typically 24 or 48) and f is the number of features retained after application of time series representation techniques and then by a factor of n^2, where n is the number of customers via application of the k-medoids clustering, with automatic criterion to each customer separately.

2.3. Block 3: Distance to Wind

Since the purpose of DR in the context of high renewables penetration is to have the load following renewable generation patterns as much as possible, for a given customer, the closest his or her load shape is to the wind generation shape, the more valuable would be any DR measure applied in this customer. This concept has been applied in this paper by computing the dynamic time warping (DTW) distance between the load shape of each customer and the wind speed load shape [40]. Two considerations are important to be mentioned in this regard. First, wind speed at a representative point has been used as a proxy of wind generation potential. It is well known that the relationship between wind speed and wind generation is not linear, but it follows the so-called power curve model [41]. In this model, a wind turbine would only start generating after a certain threshold of wind speed has been reached (cut-in speed, typically around ~3 m/s), then turbine output increases roughly linearly with wind speed until the rating of the turbine is reached (rated speed, at around ~15 m/s) and further increases in speed do not increase power generation. Finally, the turbine stops generation when a certain limit of maximum speed is reached (cut-out speed, typically around ~25 m/s). However, computing the total power output for a large set of turbines is a cumbersome exercise. On one hand, each turbine model would have a different power curve and, on the other hand, a turbine power model should have as an input the wind speed at the precise coordinates and height of each turbine to be accurate. Hence, a simplified approximation is used in this paper where the wind speed shape for a representative location is used as a proxy for wind generation pattern. Secondly, the computation of DTW in time series is an optimization problem that needs large computational resources and time to be executed. A simplified approach is proposed where only the distance between load shape medoids of each customer and the wind speed shape medoids is computed. Therefore, the total distance between a customer load shape and the wind speed shape would be a linear function of the distance between load shape and wind speed shape medoids, computed using the same 2-step k-medoids partitioning approach described above, and applied to the normalized load shape of each customer.

2.4. Block 4: Customer Features and DR Applications

Each customer is finally represented by a set of four easy-to-interpret features, two of them based on the vector of energy use (i.e., daily mean of hourly energy use and energy-temperature correlation) and two based on the normalized load shape (i.e., load shape entropy and distance to wind). These features are formally defined as follows:

- Daily mean of hourly energy use (in kWh): a large hourly energy use may be interpreted as larger potential impact in DR applications:

$$\bar{l}^i = \frac{\sum_{j=1}^{d_i} \bar{l}_j^i}{d_i} \tag{1}$$

where, $\overline{l}^i$ is the daily mean average for customer i, $\overline{l}_j^i$ is the average hourly energy for customer i and day j, and d_i is the number of days available in the time series for customer i, typically 365.

- Energy–temperature correlation (Pearson correlation factor between daily mean energy and mean temperature): a strong negative correlation may indicate a larger capability to control temperature-sensitive equipment and therefore have short-term impact on DR applications:

$$\rho_{\overline{E}_i,\overline{T}_i=\frac{cov(\overline{E}_i,\overline{T}_i)}{\sigma_{\overline{E}_i}\sigma_{\overline{T}_i}}} \tag{2}$$

  where $\overline{E}_i$ is the vector of average daily energy and $\overline{T}_i$ the vector of average daily temperature applicable to customer i.

- Load shape entropy: larger entropy means more variability, positive for DR potential.

$$Entropy_i = -\sum_{j=1}^{m_T} f_j^i \times \log_2 f_j^i \tag{3}$$

  where f_j^i is the relative frequency of each representative k-medoid for customer i and medoid j and m_T the total number of medoids representatives resulting from the 2-step clustering.

- Distance to wind: being closer to wind generation patterns means easier impact in accommodating load shape to renewable energy source (RES) generation.

$$Dwind_i = \frac{\sum_{j=1}^{d_i} DTW\left(m_j, w_j\right)}{d_i} \tag{4}$$

  where $DTW\left(m_j, w_j\right)$ is the dynamic time warping distance between the representative load medoid corresponding to day j and customer i and the corresponding wind speed medoid for day j.

3. Results

The widely used Irish CER dataset has been chosen to illustrate the application of the proposed methodology [42]. This dataset has three main advantages: (1) it is publicly available and therefore results can be replicated; (2) the data are of high quality and significant size; and (3) it contains a detailed customer survey that can be used as ground truth. Furthermore, weather data can be easily obtained from the Irish Meteorological Service historical data service [43]. This section is organized according to the methodology described in the previous section, addressing each of the blocks in turn.

3.1. Block 1: Data Analysis

The Irish CER smart meter trial contains three different tables: the half-hourly consumption data per customer for residential, SME and other customers, the allocation file, containing customer metadata (i.e., type of customer, tariff, trial group codes) and a pre-trial survey where each customer provides answers to questions related to socioeconomic profile, premises type, equipment, and attitude towards energy use. After data extraction and preprocessing, the half-hourly consumption file was filtered to retain 4224 customers containing records for the year 2010. Next, the filtered dataset was treated for missing values. The data quality is very high: ~86% of the customers has a complete 2010 half-hourly time series energy use measurement, ~13% miss only one day worth of data, and ~1% are missing two or three days. The time series of the ~14% of customers with three or less days missing has been completed using Seasonally Splitted Missing Value Imputation [44]. Then, as described in the methodology section, the energy consumption dataset has been split between amount of energy used, measured by daily half-hourly average usage and energy use shape, and is represented by the half-hourly times series normalized by the daily energy mean.

Figure 3 shows the histogram of the half-hourly daily mean average values and their range by quarter and day of the week. The histogram of half-hourly daily mean averages shows a clear unimodal distribution with no signs of differentiated segments of customers. The visual analysis of the range per day of week and quarter indicates that there is no significant difference in total energy use between days of the week and that there is clear difference between the fall/winter period (quarters 1 and 4) and the spring/summer period (quarters 2 and 3).

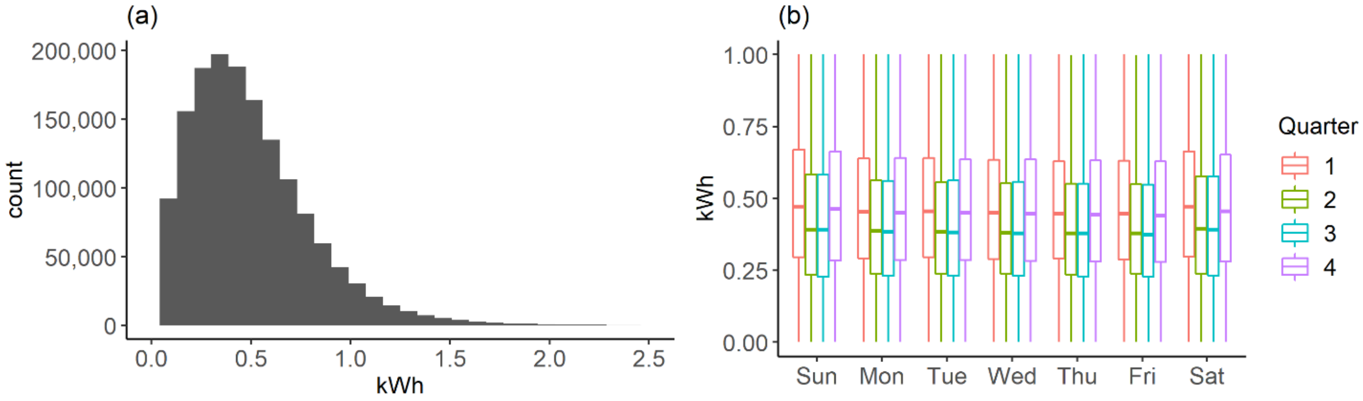

Figure 3. (**a**) Half-hourly daily mean average distribution. (**b**) Half-hourly daily mean average range by quarter and day of the week.

Figure 4 shows average values for the whole dataset of daily-normalized energy use shape by both day of the week and quarter of the year. Two clusters or typical days shape emerge from each of the figures, differentiated by the hour and intensity of the morning, afternoon and evening peaks, and driven by changing customer habits during weekend/weekdays and summer/winter months. Weekends show a clear different pattern than weekdays, with a later and larger morning peak and a relatively lower evening peak, which starts changing trend by Friday. Fall/winter months show a later and slightly shorter morning peak and a much larger and earlier evening peak.

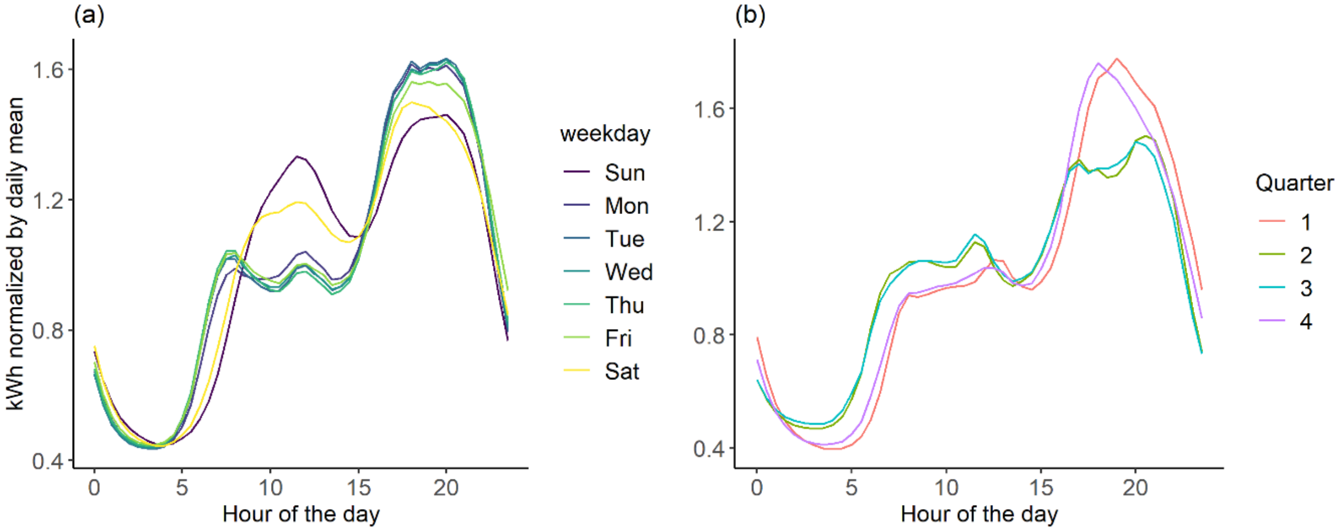

Figure 4. (**a**) Daily normalized energy use shape by day of the week. (**b**) Daily normalized energy use shape by quarter of the year.

Irish CER smart meters trial conducted surveys before and after the deployment of the meters to characterize the customer base of the trial and assess its impact on attitudes towards different energy efficiency and conservation initiatives. The pretrial survey contains a total of 143 questions covering aspects such as dwelling features, electrical equipment for heating and cooking, number of electrical and entertainment appliances and attitudes towards energy use. Our research focuses on the impact of electrical equipment in energy use and shape patterns. The existence of specific equipment is particularly important in assessing the potential to participate in DR programs by a given customer. Therefore, the answers to five questions have been extracted and combined with energy use data, covering different types of equipment: electrical central heating, plug-in heaters, heat water immersion, heat water instantaneous heaters and cookers (for all answers, "1" means that the customers have the electrical equipment and "0" that they do not have it or is a different energy source). The combined dataset of half-hourly energy use and electrical equipment existence from survey has a total of 3487 customers with the following percentages of customers having a specific electrical equipment according to the survey results: electrical central heating: 4.2%; plug-in heaters: 3.5%; heat water immersion: 55.9%, heat water instantaneous heater: 1.5% and electrical cooker: 69.7%. Therefore, most of the customers use another type of energy source for heating but, on the other hand, many customers use electricity for cooking and water heating of the immersion type. Figure 5 illustrates the impact of electrical equipment ownership on both daily energy use and demand shape pattern for the most present appliances: water heating immersion and electrical cookers. The impact on both energy metrics is distinct but not drastic. Ownership of electrical appliances for water heating and cooking increases the total energy consumed on average, but the large overlapping in energy density curves implies that energy use cannot by itself segment the customer according to their electrical equipment ownership. A similar conclusion can be drawn by analyzing the impact of electrical equipment ownership on demand shape patterns: although the electrical equipment, the cooker in particular, do have a distinct impact on the intensity and the time of morning, afternoon and evening peaks, the magnitude of this impact is relatively mild.

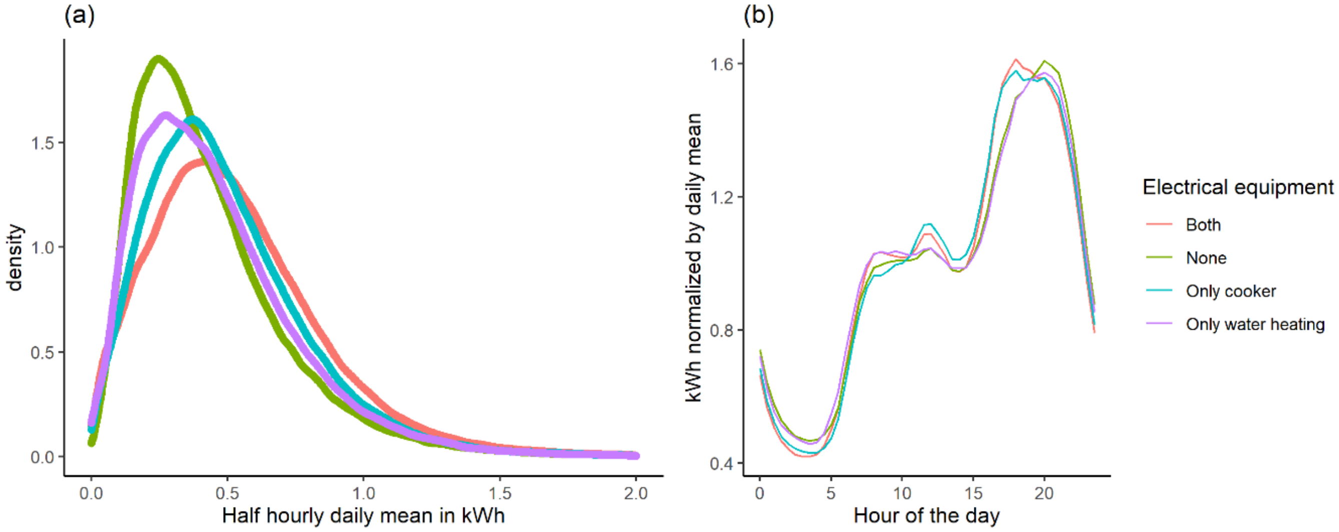

Figure 5. Impact of electrical equipment as declared in pre-trial survey on energy use and shape. (**a**) Half-hourly daily mean density plot. (**b**) Half-hourly daily normalized energy use shape pattern.

To account for the impact of weather conditions on energy demand, historical data of hourly temperature and wind speed has been extracted from the Irish Meteorological Service historical data service [43]. Since the customer metadata of the Irish CER smart meter dataset does not contain the geolocation of customers under trial, the Dublin airport weather station has been chosen as the reference for weather conditions for all customers in the dataset. Figure 6 illustrates the effect of temperature in mean daily energy use and compares the shape pattern of hourly energy demand and wind speed. Figure 6a plots the mean hourly demand per day against mean daily temperature and fits a LOESS model curve to the data, showing a typical winter-side U curve with increasing impact of the temperature starting at ~15 °C and no summer effect. The observation of Figures 4 and 5 seems to indicate that the salient differences in load shape between winter/fall and summer/spring seasons may have more to do with customer habits than the impact of electrical equipment.

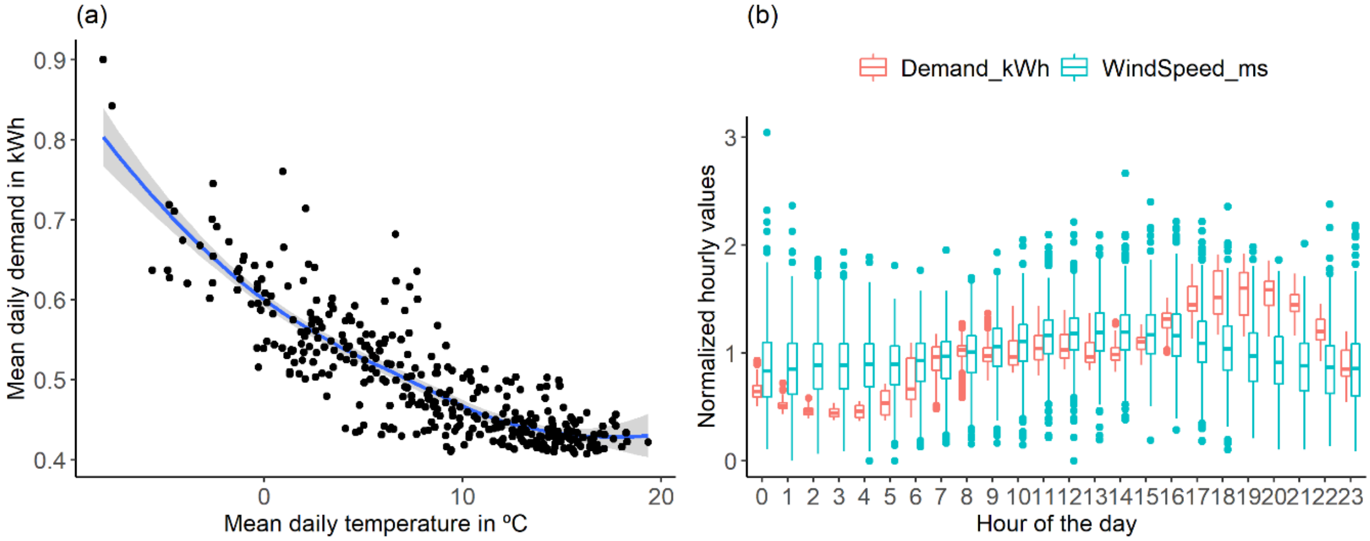

Figure 6. Weather conditions in relation to energy demand and shape patterns. (**a**) Mean hourly daily demand for all customers in kWh versus mean daily temperature. (**b**) Daily mean normalized hourly values of hourly energy demand and hourly wind speed for each hour of the day.

Figure 5b illustrates a paradigmatic difference in the variability between load shape and wind speed shape: whereas average wind speed varies within a closer range than mean load shape, total variability for each hour is much larger in wind speed than load, the former oscillating between 0 and ~3 units of normalized hourly values. Furthermore, it is relevant to note that, on average, wind generation would be close to demand in the morning and afternoon hours but lower in the evening hours and larger in the early morning. These average trends illustrate well the issue of residual demand variability (i.e., demand net of renewable generation), and, hence, the higher value of DR in the context of high penetration of wind resources.

3.2. Block 2: Two-Stage Clustering

In the first clustering step, each customer time series is broken down in daily time series and the dimensionality of each day is reduced by applying and comparing three times series representation techniques: PAA, SAX and clipping. To illustrate the application of the methodology, Figures 7 and 8 show the application of these techniques to the load curve of one customer of the dataset, ID 1260, for 1 January 2010.

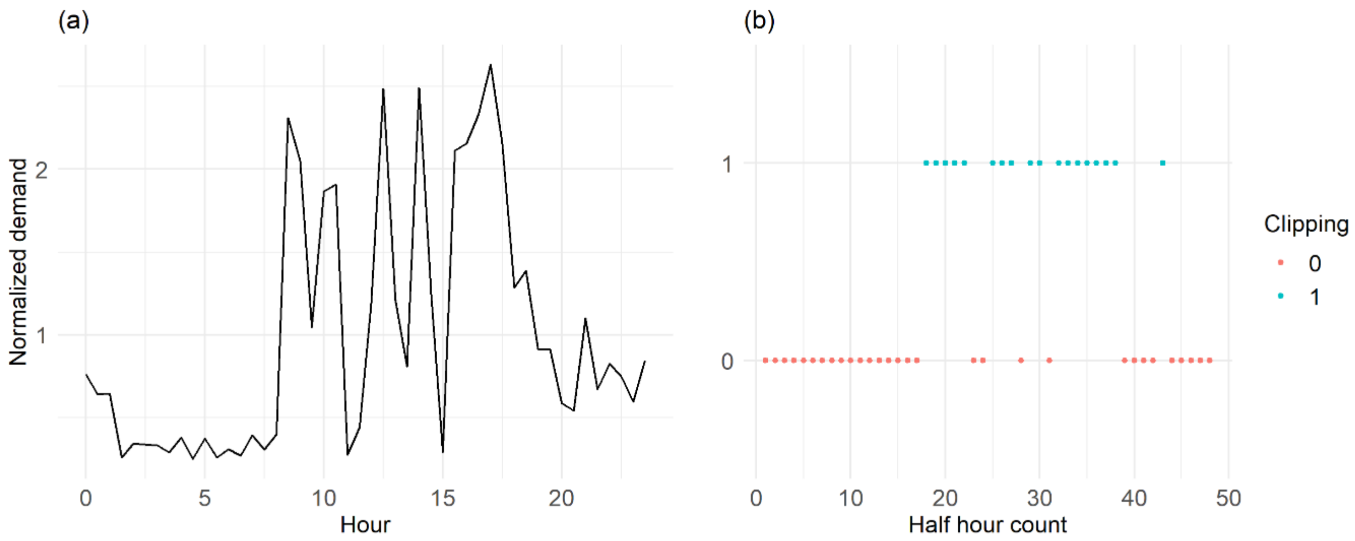

Figure 7. Application of clipping time series representation of a typical day of normalized demand. (**a**) Original mean daily normalized half-hourly demand for customer ID 1260 for 1 January 2010. (**b**) Clipping representation of the load curve for that day.

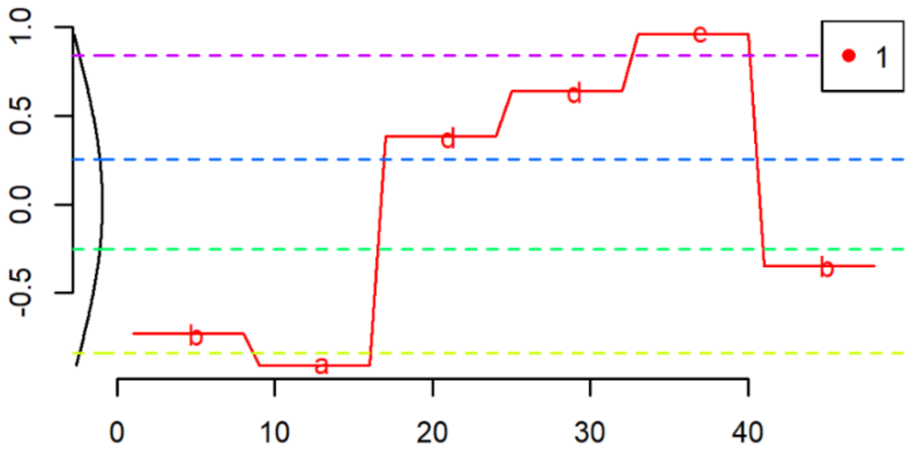

Figure 8. Application of PAA and SAX time series representation to the scaled mean daily normalized half-hourly demand for customer ID 1260 for 1 January 2010.

Given that the average mean daily normalized consumption over one day is equal to 1, the application of the clipping technique is straightforward: normalized half-hourly consumption over 1 receives a 1 mark and those below 1 receive a 0 mark. This customer day is then represented by eight features built from the clipping series: max. from run lengths of ones (7 in this case), sum of run lengths of ones (18 in this case), max. from run lengths of zeros (17 in this case), crossings (or length of run length encoding (RLE) encoding minus one, 10 in this case), number of first zeros (17 in this case), number of last zeros (5 in this case), number of first ones (0 in this case) and number of last ones (0 in this case). These easy-to-interpret features are then used to cluster customer-days according to similarity. The PAA and SAX representations need the definition of two parameters: the number of subdivisions for the day and the number of levels of representation of the demand. In our case, each day is divided into six periods (4 h each) and five levels of representation (letters "a" to "e"). After scaling the data to a standard normal distribution, the customer-day time series is represented by the average for each period in the PAA representation and by the corresponding letter in SAX.

Comparison of results of the first clustering phase with automatic stop criterion for the whole dataset of 4224 customer, and the three time series representation techniques is summarized in Table 1. Features on clipping-based reduction produces considerably better results than PAA and SAX both in terms of average silhouette values, percentage of negative silhouette and computation time. Computation time for the clipping technique is roughly half of the time for PAA and SAX. It is interesting to note the efficiency of the methodology in terms of computation time: the clipping technique produces reasonably internal clustering validity metrics with an average computation time of ~0.05 s per customer (i.e., 240 s for 4224 customers). In terms of average number of clusters, clipping has slighter lower mean and standard deviation than PAA and SAX, but in the same order of magnitude of mean ~ 2.5 clusters per customer, and a standard deviation of ~1. SAX offers the best representativity in terms of number of clusters, but with a significantly lower mean average silhouette and higher percentage of negative silhouette values.

Table 1. Comparative results of first clustering for three dimensionality reduction techniques.

Technique	Time to Reduce and Cluster (s)	Mean of Average Silhouette	Mean of % of Negative Silhouette	Mean of Number of Clusters	Standard Dev. Number of Clusters
PAA	536	0.29	5.91	2.49	0.98
SAX	629	0.18	9.98	2.63	1.08
Clipping	240	0.35	3.42	2.46	0.76

In the second clustering phase, the medoids of each customer are extracted and a new k-medoids clustering is undertaken in the population of 10,560 customer medoids, after the representative time series medoids have been reduced using the same techniques as in phase one of clustering. Figure 9 compares the average silhouette values for each dimensionality reduction technique and different number of clusters. Again, clipping shows a clearly better clustering performance than the PAA and SAX alternatives. Average silhouette reaches a plateau of ~0.4 for 3–4 clusters for clipping reduced representative medoids to drop and stabilize at around ~0.3 for five clusters and above. Both PAA and SAX stabilize at an average silhouette of ~0.2 for number of clusters of four and beyond, with a peak at ~0.3, and only two clusters in the case of PAA. In summary, clipping provides acceptable clustering results for both phases in terms of internal validation metrics and a better alternative than PAA and SAX as a dimensionality reduction technique.

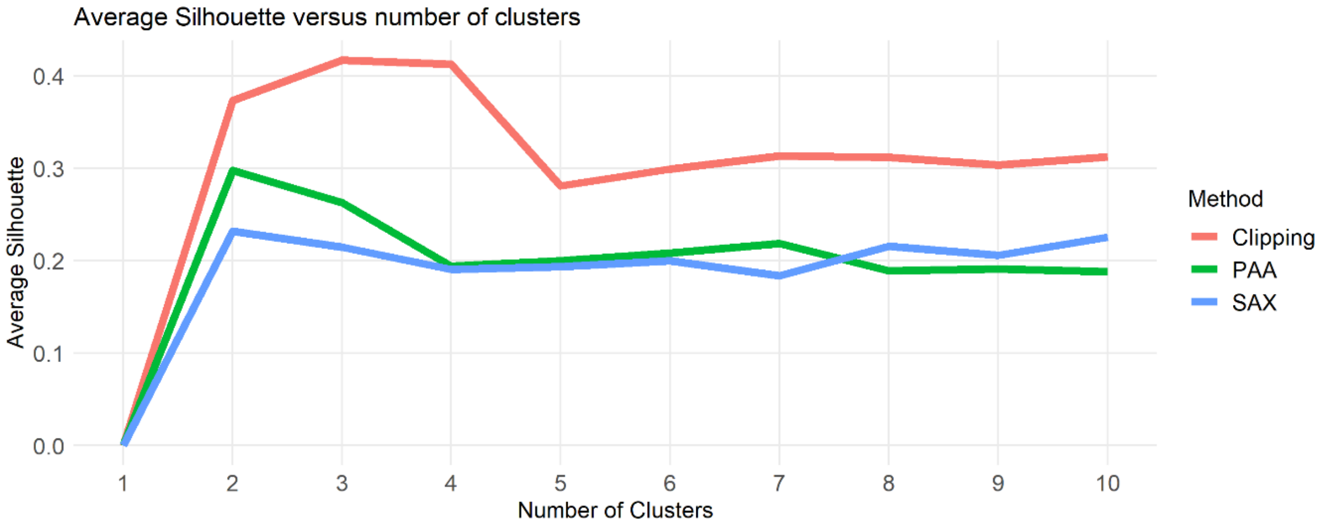

Figure 9. Second clustering phase average silhouette versus number of clusters per technique.

In addition to the comparison between time series representation techniques in terms of internal validation metrics, an external validation comparison has been undertaken using survey data as a ground truth. Each customer is represented by the relative frequency of each cluster and this feature is used to build univariate logistic regression models to predict the probability of existence of a given piece of equipment in each customer. In mathematical notation, $logit(p) = log\frac{p}{1-p} = \beta_0 + \beta_i \times f_i$, where p is the probability of a given customer to have the equipment, f_i the relative frequency of cluster i, and β_i the corresponding logistic regression coefficient. Table 2 shows the positive regression coefficients for each paired equipment–technique model, which are statistically representative with a p-value < 0.005. For all techniques, the number of clusters has been set to four in order to make the regression coefficients comparable between techniques. The interpretation of the positive logistic regression coefficients with p-value below 0.05 is that a higher proportion of cluster i in a given customer means a higher probability of the customer having this piece of equipment.

Table 2. Comparative results for equipment detection using logistic regression coefficients.

Equipment	PAA	SAX	Clipping
HeatingE_central	-	-	-
HeatingE_plugin	$\beta_2 = 0.88$	$\beta_1 = 1.05$	$\beta_3 = 1.95$
Water_HeatingE_inmersion	-	$\beta_3 = 0.31$	-
Water_HeatingE_instant	-	-	-
Cooker_type	$\beta_4 = 0.92$	$\beta_3 = 0.93$	$\beta_2 = 0.75$

Three main insights can be drawn from Table 2. Firstly, the clustering methodology is effective in detecting the presence of certain equipment in the final clients, as positive regression coefficients with high statistical representativity are found for the three representation techniques. Secondly, only the equipment that has a distinctive impact on normalized load shapes can be captured by the regression model. In this case, only the plug-in electrical heating and the electrical cooker seem to have a relatively sufficient impact on load curves to be detected. Thirdly, the results among the three techniques are quite similar and consistent, detecting the same pieces of equipment (with the exemption of SAX being the only one to detect the water heating immersion, albeit with a small value of regression coefficient) and with absolute values of regression coefficient of a similar order of magnitude. Again, the clipping technique is faring well in the comparison, detecting both the electrical heating and the electrical cooker, the first one with a regression coefficient value that is double of the SAX and PAA ones.

3.3. Block 3: Computation of Distance to Wind

The next step of the methodology, the computation of the distance between load medoids representatives and wind speed medoids representatives, is undertaken using the features clipping methodology for dimensionality reduction for both the load and wind speed shape. Figure 10a shows the four medoids that represent the full customer-day population resulting from the k-medoids clustering to the set of customer representatives. As expected, each cluster represents customer-days that have energy use peaks at different times of the day and with different intensity. Cluster 1, with ~19% of the customer representatives, shows a moderate peak in the early morning and a large peak in the evening. Cluster 2, with ~55% of representatives, is marked by a relative higher peak in the early afternoon. Cluster 3, with ~6% of representatives, has the flattest load shape profile, with a relative higher peak in the early morning. Finally, cluster 4, with ~20% of representatives, shows three distinctive peaks of a similar order of magnitude in the early morning, late morning and late afternoon. Similarly, Figure 10b shows the profile of the wind speed time series representatives k-medoids clustering results. The k-medoids clustering of the daily normalized wind speed time series generates three clusters with an average silhouette of 0.26% and 5.5% of negative silhouette values. Each cluster represents a wind speed daily pattern. Cluster 1, representing ~40% of the days, has the relative peak early afternoon. Cluster 2, representing ~52% of the days, has a similar peak early afternoon, but higher speeds during night hours. Finally, cluster 3, with 8% of representation, shows the relative peak in the early morning and decreases until the later hours of the night.

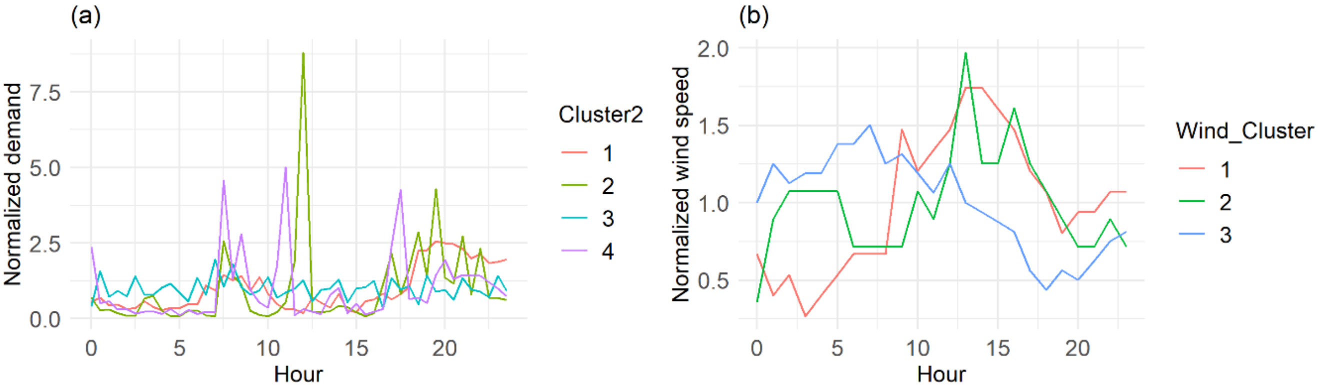

Figure 10. Extraction of load shape and wind medoids representatives. (**a**) Four medoid representatives for load shape. (**b**) Three medoid representatives for wind speed shape.

The distance DTW distance algorithm, which computes the distance between two time series by stretching/compressing them locally to make them as similar as possible, has been used to compute the similarity matrix between the set of load shape and wind speed medoids representatives. Figure 11a shows as an example of the alignment between the medoid representing load shape cluster 1 (query or test) and the medoid representing wind speed cluster 1 (reference). The distance between the two time series is equal to the sum of the (unnormalized) Euclidean distance between the aligned points of each time series.

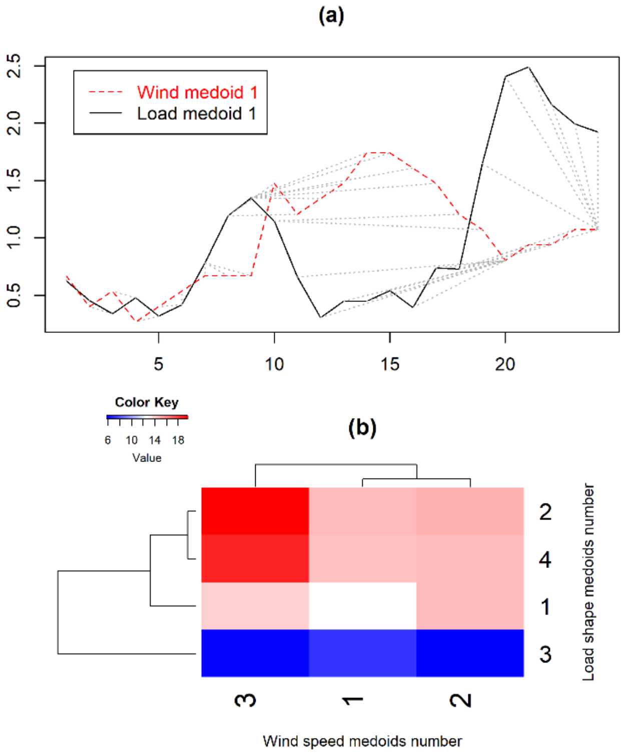

Figure 11. Computation of distance matrix between load and wind speed medoids using DTW dissimilarity. (**a**) Illustration of computation of DTW distance for one combination load-wind speed medoids. (**b**) Distance matrix in heatmap format.

The distance between load and wind speed shape patterns is a per unit metric, as both load and wind shape are per unit, daily-mean normalized values, which by definition fulfill the following condition:

$$\sum_{h=1}^{24} \bar{l}_{i,h} = \sum_{h=1}^{24} \overline{w}_{j,h} = 24 \tag{5}$$

where $\bar{l}_i$ is the i-th load medoid vector, and $\overline{w}_j$ is the j-th wind speed medoid vector, both in a 24-h format. Figure 10b shows the DTW distance between each of the load shape and wind speed medoids in the form of a heatmap. This type of representation allows to graphically assess the relative distance of the different load patterns as defined by the k-medoid clustering algorithm and the wind speed patterns. For instance, it can be noted that load shape cluster 2, the more representative of the population with ~55%, is the cluster with a larger distance to the three wind speed k-medoid representatives, whereas load shape cluster 3, representative of ~6% of customer-days, is the closest to wind speed patterns. The heatmap representation allows also for a relative distance graphical assessment among the wind speed and load shape medoids, as illustrated by the row and column dendrograms. Load shape cluster 3 and wind speed cluster 3 are the most distinctive representatives when compared to the other clusters of their same group.

3.4. Block 4: Customer Features and DR Applications

Each customer is then represented by four features, two derived from the energy consumed (daily average of hourly energy use and daily correlation with temperature) and two derived from the shape of this consumption (entropy of medoids representatives and distance to wind). Additionally, the load shape of each customer may also be represented by the relative frequency of each of the k-medoid representatives of the second phase of clustering. This feature's representation can then be used for different applications related to the selection of customers for DR programs by utilities. In this section, three applications are illustrated: visual representation of DR potential per customer, clustering of customers according to DR potential and detection of electrical equipment.

Figure 12 shows the first application; the visual representation of the DR potential in one single graph using the four features. This visual representation is inspired by [33], where DR potential is visualized in a 2-D representation with a measure of flexibility (entropy of cluster vector) in the x-axis, and energy use in the y-axis, both variables in percentiles. Customers with higher potential would be in the upper-right quadrant of the figure. Similarly, a 2-D representation is used in our case where customers with higher DR potential are in the upper-right quadrant and energy use is also represented in the y-axis. However, our representation uses an inverse scale of the distance to wind in the x-axis, meaning that customers with higher total consumption and closest to wind patterns in shape would have greater DR potential. The graphic representation is further enriched by showcasing both the correlation to temperature with color and entropy with size for each customer, in such a way that customer prioritization can be further undertaken for a given energy vs. distance-to-wind quadrant. Furthermore, both absolute values and percentiles are represented to facilitate the assessment of DR potential for each customer.

The second application, customer segmentation according to DR potential, is undertaken using a simple but effective k-means clustering algorithm. After min-max normalization of the four features, the number of clusters is determined using both the elbow and gaps-stats methods, resulting on an optimal number of five clusters. The ratio of between-clusters sum-of-squares and total sum-of-squares is 65%, suggesting an acceptable clustering quality. From the visual representation of the density function for each feature and cluster shown in Figure 13, a characterization of the five clusters according to DR potential can be attempted. The selection of segments for DR programs would depend on the specific feature to be prioritized. If distance to wind is to be emphasized, cluster 5 (173 customers, ~4% of population) is the only segment differentiating with a significant lower distance to wind. However, it shows the lowest hourly energy use and average values for entropy and temperature correlation. If flexibility (measured as higher entropy) is to be highlighted, cluster 2 (584 customers, ~14% of population) offers a clear differentiation, together with relative larger energy use and average temperature correlation. Cluster 4 (1253 customers, ~30% of population) offers the largest potential in the combination of energy use and correlation to temperature together with average flexibility. Clusters 1 (1047 customers, ~25% of population) and 3 (1167 customers, ~28% of population) are the ones that are less interesting for DR applications, having average energy use values while showing very low flexibility in the case of cluster 3, and low correlation with temperature in the case of cluster 1.

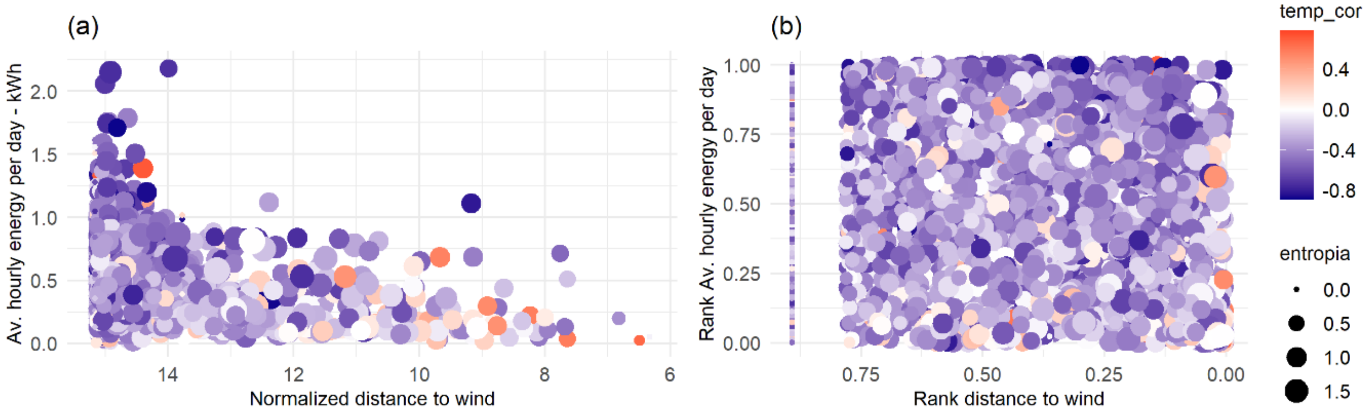

Figure 12. Visualization of customers' DR potential according to features. (**a**) Average hourly energy and distance to wind in absolute values. (**b**) Same representation in percentiles.

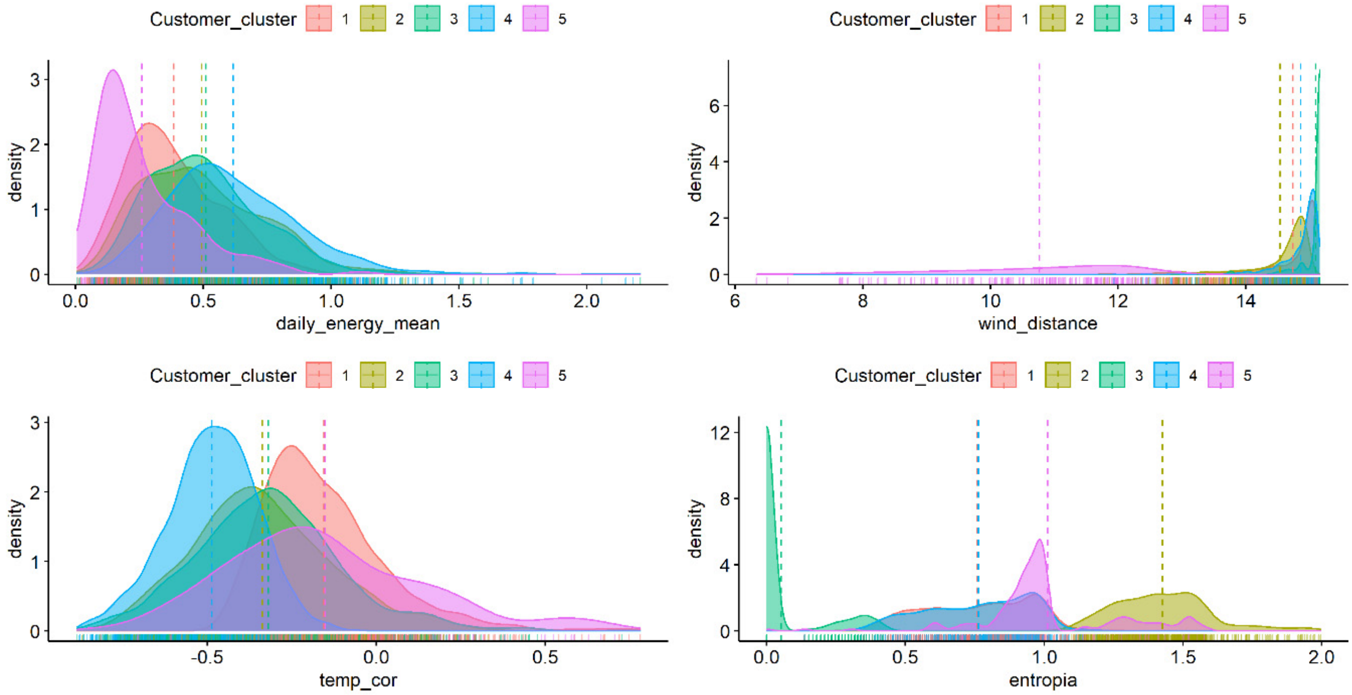

Figure 13. Density function for each customer feature and cluster.

The third application is the detection of electrical equipment using the relative frequency of each load-shape cluster. A similar approach to that followed in the external validation of the second phase of load shape clustering, summarized in Table 2, is carried out. Table 3 completes the analysis for the two most significant pieces of equipment, plug-in electrical heating and cooker type, adding mean hourly energy and correlation to temperature as explicative variables. Both positive and negative beta coefficients are shown, but only when they are statistically significant (i.e., corresponding p-values below 0.005). The probability of a customer having a plug-in electrical heater is larger when the correlation to decreasing temperatures is higher and when the proportion of cluster 3 is higher. Similarly, the probability of having an electrical cooker is higher when the energy use and the proportion of cluster 2 are larger.

Table 3. β logistic regression coefficients for univariate models and p-value < 0.005.

Equipment	Hourly_Energy_Mean	Temp_Cor	Cluster_Pro1	Cluster_Pro2	Cluster_Pro3	Cluster_Pro4
HeatingE_plugin	-	−3.06	-	-	1.95	-
Cooker_type	1.28	-	−1.48	0.75	-	−0.42

To assess the prediction power of the generalized linear model, a generalized linear model is defined by combining the effect of the two statistically relevant explanatory variables on the probability for a given customer to have the electrical equipment target:

$$log \frac{p_{plug-in}}{1 - p_{plug-in}} = \beta_0 + \beta_1 \times temp\_cor + \beta_2 \times cluster\_pro3 \tag{6}$$

The results of the two-variable generalized linear model show that both explanatory variables are statistically relevant (p-values well below 0.001) with $\beta_1 = -3.26$ and $\beta_2 = 2.74$. However, the total explanatory power of the model is weak. Residual deviance is ~94% of total deviance, meaning that only ~6% of the variance of the target variable can be explained by the two explanatory variables. Therefore, it can be concluded that, for this specific dataset, customers with a higher negative correlation with temperature and a higher proportion of customer-day k-medoid number 3 are more likely to have electrical heating plug-in equipment, but the explanatory power of these two variables is not good enough to build a reliable prediction model. This result can be explained by the relatively weak impact of electrical equipment in normalized load patterns, as illustrated in Figure 5.

4. Discussion

The proposed methodology and its application on the Irish smart meters trial dataset contribute to new knowledge along three main axes: the effectiveness of the times series clustering approach, the definition of customer features representation adapted to DR in the context of high penetration of renewable energy, and the design of DR applications for customer selection.

Time complexity reduction has been achieved by a combination of two techniques: dimensionality reduction and a two-stage clustering approach. The use of the least computational demanding dimensionality reduction techniques, such as PAA, SAX and features-based clipping, allow for an effective application of the methodology in a relatively large dataset using a laptop. Specifically, clipping produces better internal validity clustering scores, while taking half the time of the rest of the techniques. Hence, feature-based clipping has been used, for the first time, to the knowledge of the authors, end-to-end to cluster the normalized smart meters time series. Similarly, the first of the two-stage k-medoids clustering approach, i.e., the application of the PAM algorithm to the reduced time series of each customer separately, allows for a drastic reduction in the time complexity by a factor proportional to the square of the number of customers. This combined approach reduces the 48 time-dimensions of each customer-day to eight features, hence a reduction factor of six, and undertakes the first phase clustering of the full dataset in only ~0.05 s per customer. The effectiveness in managing time complexity does not preclude a performant clustering result as measured by the internal validity metric of average silhouette and the external validation metric using the ground truth contained in surveys. The features-based clipping approach achieved an average silhouette of ~0.4 in the second clustering phase, with less than 1% of negative silhouette values and proving to be effective in detecting electrical equipment from survey replies using a logistic regression approach. The two-stage clustering on a customer-days matrix is similar to the adaptive k-means implemented in [21] and the LSD approach in [32]. Yet, it is also different on several accounts: by using an automatic stop criterion to determine the number of clusters for each customer without the definition of a threshold, and by using medoids as centroids instead of k-means, as in previous work [32,33].

The set of customer features defined, two relating to absolute energy use and two relating to load-shape patterns, provides an innovative framework for assessing customer's DR potential. Daily energy use, or similar metrics accounting for the absolute value of energy use regardless of time-shape patterns, has been widely used as a representative customer feature, often broken down in typical periods of the day (i.e., night/morning/afternoon/evening) [20,21,45,46]. The correlation between energy use and temperature has also been extensively used in the literature [18,19,21]. Entropy as a measure of variability has also been previously used [16,27,30]. In [33], the segmentation of customers regarding potential for DR is carried out in terms of quantity—total daily energy—and variability, measured as entropy of the number of encoded load shapes per customer. The key contribution of the methodology is taking explicitly into account the impact of renewable energy generation patterns. DR potential has often been measured in terms of contribution to the system peak for specific customers [47,48]. Instead, our proposal is to use a metric better adapted to the needs of systems with high penetration of renewables, distance to wind, measuring the similarity between the customer load pattern and wind speed, a proxy for wind generation patterns. The approach has been proven to be successful using the DTW distance, overcoming the large computing demands of DTW distance by applying the raw data of only the medoids of both customer-day load profiles and wind speed daily curves.

These four customer features, derived both from the amount of energy used and its load-shape pattern, are the base for designing different applications related to DR program implementation in the context of a high penetration of renewable resources. In this paper we have illustrated three of them: the visualization of the customers in 2-D space to select candidates for DR programs (inspired by [33] but adding new variables and visual features), the clustering of customers according to their potential for DR applications using k-means and, for the first time, to the knowledge of the authors, the use of frequency of representatives as a predictor of energy equipment using logistic regression.

5. Conclusions

This paper has successfully addressed the two-fold research goal of designing and implementing a computationally efficient smart meters time series clustering methodology that takes explicitly into account renewable energy generation patterns. The two-stage clustering approach is also scalable to larger datasets. Transforming the whole dataset of customer-days into a linear combination of each customer's dataset makes the first stage of clustering scalable to much larger datasets than a few thousand customers. Additionally, several additional techniques could be used to make the second clustering phase scalable, such as the application fast k-medoids algorithms CLARA or CLARANS for large datasets [49], or the use of distributed computing techniques [50]. Further research could test the scalability of the methodology in datasets one order of magnitude larger than the one used in this paper (+100,000 customers with at least one year's worth of hourly smart meters data).

A novelty of this paper that could also be further developed in future research is the use of time series clustering features as predictors of ground truth embedded in customer surveys. Whereas the combination of weather data with energy use profiles is relatively common in the literature, the simultaneous use of smart meters time series data, locational weather and customer surveys is not. The use of logistic regression to determine the likelihood of the existence of electrical equipment using time series representative's frequency as predictors could be extended to other customer features such as socioeconomic status, type of premises or attitude towards energy conservation. Furthermore, it would be insightful to apply this methodology to other smart meters datasets containing customers with a more sophisticated use of equipment such as electrical vehicles, storage, heat pumps or distributed solar [16].

An additional line of further research would be the adaptation of the methodology to the actual structure of demand flexibility services as they are being defined and implemented in advanced electricity markets [51]. Standard dimensionality reduction techniques such as clipping, SAX or PAA could be compared with the definition of expert-based measures using developing flexibility metrics such as equivalent ramping capability (ERC) or ramping availability rate (RAR). A far-fetching research approach would be to address the short-term nature of ramp or energy flexibility products. Whereas the current approach of considering a one-year long time series dataset may be well suited for structural, long-term flexibility products such as capacity markets, a data streaming approach would be necessary to ascertain the DR potential and value in real time markets (e.g., 5 min frequency), intra-day (e.g., 1 h frequency) or daily (e.g., 24 h ahead) flexibility markets.

Author Contributions: Conceptualization, S.B.; methodology, S.B.; supervision, R.D. and N.D.; writing—original draft, S.B.; writing—review and editing, R.D. and N.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Irish smart meter trial CER dataset has been accessed via the Irish Social Science Data Archive—www.ucd.ie/issda (accessed 26 June 2020) in accordance with the terms of use of the license agreement. Copyright and all other intellectual property rights in the data and associated documentation are vested in the original data creators or depositors. The authors of this paper acknowledge the work of the original data creators, depositors and copyright holders and the ISSDA, and declare that those who carried out the original analysis and collection of the data bear no responsibility for the further analysis or interpretation of it. Copyright and all other intellectual property rights in the data and associated documentation are vested in the original data creators or depositors. Weather data have been obtained from the Irish Meteorological Service (Met Éireann) historical data service (source www.met.ie) (accessed 26 June 2020). Copyright Met Éireann. These data are published under a Creative Commons Attribution 4.0 International (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/ (accessed 26 June 2020). Disclaimer: Met Éireann does not accept any liability whatsoever for any error or omission in the data, their availability, or for any loss or damage arising from their use.

Acknowledgments: This work was supported in part by the Spanish Ministry of Economy and Competitiveness under the Project CICYT RTI2018-094665-B-I00, the Project PID2019-108377RB-C32 and the Project DPI2017-84259-C2-2-R. In addition, the Project GID2016-6 supported by UNED.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

AI	Artificial Intelligence
AMI	Advanced Metering Infrastructure
DMS	Demand Side Management
DR	Demand Response
DTW	Dynamic Time Warping
HC	Hierarchical Clustering
LSD	Load Shape Dictionary
ML	Machine Learning
PAA	Piecewise Aggregate Approximation
PAM	Partition Around Medoids
PCA	Principal Component Analysis
SAX	Symbolic Aggregate Approximation
SOM	Self-Organized Maps
TOU	Time Of Use

Appendix A

Mathematical Notation for the Two-Step k-Medoids Clustering Approach

The original energy use dataset is organized in a matrix format where, for each customer, each row represents a day and each column a daily time division of the day:

$$y^i_j = \left\{l^i_{j,1}, \ldots,\ l^i_{j,h}\right\},\ i \in N\{1, \ldots,\ n\},\ j \in N\{1, \ldots, d_i\}$$

where $n \in N$ is the number of customers in the dataset, $d_i \in N$ is the number of days with smart meter data for customer i (typically 365, one whole year), and $h \in N$ is the number of time periods per day (i.e., 24 for hourly data, 48 for half-hourly and 96 for $15'$ frequency data).

As a second step, each load-shape for customer i and day j is normalized by dividing each unit of energy use by the average of the day:

$$\bar{y}^i_j = \{\frac{l^i_{j,1}}{\bar{l}^i_j}, \ldots, \frac{l^i_{j,h}}{\bar{l}^i_j}\}$$

where $\bar{l}^i_j = \frac{\sum_{p=1}^{h} l^i_{j,p}}{h}$, is the mean daily energy for customer i and day j.

Algorithm A1 is applied to each customer normalized load shape to obtain a transformed dataset and to compute the distance matrix between the customer-days elements.

Algorithm A1 Time Series Transformation and Distance Matrix Computation

```
Require: Normalized load-shape time series for customer s, $\bar{y}^s = \left\{\bar{y}^s_1, \ldots, \bar{y}^s_{d_s}\right\}^T \in N^{d_s \times h}$
Set clock
Perform time series representation transformation
$f(\bar{y}^s) \rightarrow \bar{t}^s = \left\{\bar{t}^s_1, \ldots, \bar{t}^s_{d_s}\right\}^T \in N^{d_s \times f}, f < h$
Min-max normalize of time series representation features
Compute distance matrix for $d_s$ customer-days normalized representation
Stop clock
Compute time_to_distance
Return time_to_distance, distance_matrix
```

Algorithm A2, a k-medoids clustering with automatic stop criteria, is applied to the customer-day elements of each customer in the dataset. The stop criterion is based on the computation of the average silhouette for each iteration of increasing number of clusters: the algorithm stops when the average silhouette decreases with the number of clusters. The medoids of each customer-day clustering are the representatives of each one of the customers. The sample size is radically reduced from $\sum_{i=1}^{n} d_i$ to $\sum_{i=1}^{n} m_i$, where m_i is the number of medoids for customer i and $m_i \ll d_i$.

Algorithm A2 k-Medoid Clustering with Automatic Stop Criterion for Single Customer

```
Require: Distance matrix for $d_s$ customer-day items for customer $s$, Max number of clusters (max.k)
Set clock
Compute k-medoids clustering with 1 cluster
Set average.silhouette[1 cluster] equal to 0
For k in 2 to max.k do
    Compute k-medoids clustering with k clusters
    If average.silhouette[k clusters] < average.silhouette[$k - 1$ clusters] then stop
Stop clock
Compute time_to_cluster
If k = k.max then number_of_clusters C = k else number_of_clusters C = $k - 1$
Return time_to_cluster, number_of_clusters, average_silhouette[C clusters],
percentage_negative_silhouette[C clusters], medoids_index[C clusters], cluster_vector[C clusters]
```

Finally, the same k-medoids algorithm is used with the set of population representatives medoids $\left\{\bar{r}_1^1, \ldots, \bar{r}_{m_n}^n\right\}^T \in N^{(\sum_{i=1}^{n} m_i)\times f}$. As a result of this second clustering step, a reduced set of m_T medoids $\{\bar{r}_1, \ldots, \bar{r}_{m_T}\}^T \in N^{m_T \times f}$ representatives for the whole population is obtained, where $m_T \ll \sum_{i=1}^{n} m_i$.

References

1. Parrish, B.; Heptonstall, P.; Gross, R.; Sovacool, B.K. A systematic review of motivations, enablers and barriers for consumer engagement with residential demand response. *Energy Policy* **2020**, *138*, 111221. [CrossRef]
2. Bañales, S. The enabling impact of digital technologies on distributed energy resources integration. *J. Renew. Sustain. Energy* **2020**, *12*, 045301. [CrossRef]
3. International Energy Agency. *Global Energy Review 2020*; IEA: Paris, France, 2020. Available online: https://www.iea.org/reports/global-energy-review-2020 (accessed on 10 June 2021).
4. International Energy Agency. World Energy Outlook 2020–Event–IEA. 2020. Available online: https://www.iea.org/events/world-energy-outlook-2020 (accessed on 10 June 2021).
5. Hussain, M.; Gao, Y. A review of demand response in an efficient smart grid environment. *Electr. J.* **2018**, *31*, 55–63. [CrossRef]
6. Haider, H.T.; See, O.H.; Elmenreich, W. A review of residential demand response of smart grid. *Renew. Sustain. Energy Rev.* **2016**, *59*, 166–178. [CrossRef]
7. Zhang, Y.; Huang, T.; Bompard, E.F. Big data analytics in smart grids: A review. *Energy Inform.* **2018**, *1*, 8. [CrossRef]
8. Akhavan-Hejazi, H.; Mohsenian-Rad, H. Power systems big data analytics: An assessment of paradigm shift barriers and prospects. *Energy Rep.* **2018**, *4*, 91–100. [CrossRef]
9. Alahakoon, D.; Yu, X. Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey. *IEEE Trans. Ind. Inform.* **2016**, *12*, 425–436. [CrossRef]
10. Antonopoulos, I.; Robu, V.; Couraud, B.; Kirli, D.; Norbu, S.; Kiprakis, A.; Flynn, D.; Elizondo-Gonzalez, S.; Wattam, S. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renew. Sustain. Energy Rev.* **2020**, *130*, 109899. [CrossRef]
11. Wang, Y.; Chen, Q.; Hong, T.; Kang, C. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Trans. Smart Grid* **2019**, *10*, 3125–3148. [CrossRef]
12. Tureczek, A.M.; Nielsen, P.S. Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data. *Energies* **2017**, *10*, 584. [CrossRef]
13. Rajabi, A.; Eskandari, M.; Ghadi, M.J.; Li, L.; Zhang, J.; Siano, P. A comparative study of clustering techniques for electrical load pattern segmentation. *Renew. Sustain. Energy Rev.* **2020**, *120*, 109628. [CrossRef]
14. Liao, T.W. Clustering of time series data—A survey. *Pattern Recognit.* **2005**, *38*, 1857–1874. [CrossRef]
15. Aghabozorgi, S.; Shirkhorshidi, A.S.; Wah, T.Y. Time-series clustering—A decade review. *Inf. Syst.* **2015**, *53*, 16–38. [CrossRef]
16. Wang, Y.; Chen, Q.; Kang, C.; Zhang, M.; Wang, K.; Zhao, Y. Load profiling and its application to demand response: A review. *Tsinghua Sci. Technol.* **2015**, *20*, 117–129. [CrossRef]
17. Cao, H.-A.; Beckel, C.; Staake, T. Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns. In Proceedings of the IEEE IECON 2013–39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, Austria, 10–13 November 2013; pp. 4733–4738.

18. Sandels, C.; Kempe, M.; Brolin, M.; Mannikoff, A. Clustering Residential Customers with Smart Meter data using a Data Analytic Approach—External Validation and Robustness Analysis. In Proceedings of the IEEE 2019 9th International Conference on Power and Energy Systems (ICPES), Perth, Australia, 10–12 December 2019; pp. 1–6.
19. McDonald, B.; Pudney, P.; Rong, J. Pattern recognition and segmentation of smart meter data. *ANZIAM J.* **2014**, *54*, 105–150. [CrossRef]
20. Grigoras, G.; Ivanov, O.; Gavrilas, M. Customer classification and load profiling using data from Smart Meters. In Proceedings of the IEEE 12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL), Belgrade, Serbia, 25–27 November 2014; pp. 73–78.
21. Kwac, J.; Tan, C.-W.; Sintov, N.; Flora, J.; Rajagopal, R. Utility customer segmentation based on smart meter data: Empirical study. In Proceedings of the 2013 IEEE International Conference on Smart Grid Communications (SmartGridComm), Vancouver, BC, Canada, 21–24 October 2013; pp. 720–725.
22. Motlagh, O.; Foliente, G.; Grozev, G. Knowledge-Mining the Australian Smart Grid Smart City Data: A Statistical-Neural Approach to Demand-Response Analysis. In *Lecture Notes in Geoinformation and Cartography*; Springer Science and Business Media LLC: Cham, Switzerland, 2015; Volume 213, pp. 189–207.
23. Verma, A.; Asadi, A.; Yang, K.; Maitra, A.; Asgeirsson, H. Analyzing household charging patterns of Plug-in electric vehicles (PEVs): A data mining approach. *Comput. Ind. Eng.* **2019**, *128*, 964–973. [CrossRef]
24. Tureczek, A.; Nielsen, P.S.; Madsen, H. Electricity Consumption Clustering Using Smart Meter Data. *Energies* **2018**, *11*, 859. [CrossRef]
25. Ryu, S.; Choi, H.; Lee, H.; Kim, H.; Wong, V.W.S. Residential Load Profile Clustering via Deep Convolutional Autoencoder. In Proceedings of the 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Aalborg, Denmark, 29–31 October 2018; pp. 1–6.
26. Laurinec, P.; Lucká, M. Comparison of representations of time series for clustering smart meter data. In Proceedings of the World Congress on Engineering and Computer Science 2016, San Francisco, CA, USA, 19–21 October 2016; Volume 1, pp. 458–463.
27. Lin, S.; Li, F.; Tian, E.; Fu, Y.; Li, D. Clustering Load Profiles for Demand Response Applications. *IEEE Trans. Smart Grid* **2019**, *10*, 1599–1607. [CrossRef]
28. Yuan, Y.; Dehghanpour, K.; Bu, F.; Wang, Z. A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand. *IEEE Trans. Power Syst.* **2020**, *35*, 4026–4035. [CrossRef]
29. Liu, C.; Wang, X.; Huang, Y.; Liu, Y.; Li, R.; Li, Y.; Liu, J. A Moving Shape-based Robust Fuzzy K-modes Clustering Algorithm for Electricity Profiles. *Electr. Power Syst. Res.* **2020**, *187*, 106425. [CrossRef]
30. Charwand, M.; Gitizadeh, M.; Siano, P.; Chicco, G.; Moshavash, Z. Clustering of electrical load patterns and time periods using uncertainty-based multi-level amplitude thresholding. *Int. J. Electr. Power Energy Syst.* **2020**, *117*, 105624. [CrossRef]
31. Lee, E.; Kim, J.; Jang, D. Load Profile Segmentation for Effective Residential Demand Response Program: Method and Evidence from Korean Pilot Study. *Energies* **2020**, *13*, 1348. [CrossRef]
32. Liang, H.; Ma, J.; Sun, R.; Du, Y. A Data-Driven Approach for Targeting Residential Customers for Energy Efficiency Programs. *IEEE Trans. Smart Grid* **2020**, *11*, 1229–1238. [CrossRef]
33. Kwac, J.; Flora, J.; Rajagopal, R. Household Energy Consumption Segmentation Using Hourly Data. *IEEE Trans. Smart Grid* **2014**, *5*, 420–430. [CrossRef]
34. Keogh, E.J.; Chakrabarti, K.; Pazzani, M.; Mehrotra, S. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowl. Inf. Syst.* **2001**, *3*, 263–286. [CrossRef]
35. Lin, J.; Keogh, E.; Lonardi, S.; Chiu, B. A symbolic representation of time series, with implications for streaming algorithms. In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD '03, San Diego, CA, USA, 13 June 2003; pp. 2–11.
36. Laurinec, P.; Lucká, M. Interpretable multiple data streams clustering with clipped streams representation for the improvement of electricity consumption forecasting. *Data Min. Knowl. Discov.* **2019**, *33*, 413–445. [CrossRef]
37. Laurinec, P. TSrepr R package: Time Series Representations. *J. Open Source Softw.* **2018**, *3*, 577. [CrossRef]
38. Gower, J.C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **1971**, *27*, 857. [CrossRef]
39. Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [CrossRef]
40. Giorgino, T. Computing and visualizing dynamic time warping alignments in R: The dtw package. *J. Stat. Softw.* **2009**, *31*, 1–24. [CrossRef]
41. Wang, Y.; Hu, Q.; Li, L.; Foley, A.M.; Srinivasan, D. Approaches to wind power curve modeling: A review and discussion. *Renew. Sustain. Energy Rev.* **2019**, *116*, 109422. [CrossRef]
42. Commission for Energy Regulation (CER). *CER Smart Metering Project–Electricity Customer Behaviour Trial, 2009–2010*, 1st ed.; Irish Social Science Data Archive: Dublin, Ireland, 2012; SN: 0012-00.
43. Irish Meteorological Service. Historical Weather Data. Available online: https://www.met.ie/climate/available-data/historical-data (accessed on 10 June 2021).
44. Moritz, S.; Bartz-Beielstein, T. imputeTS: Time Series Missing Value Imputation in R. *R J.* **2017**, *9*, 207–218. [CrossRef]
45. Yilmaz, S.; Chambers, J.; Patel, M. Comparison of clustering approaches for domestic electricity load profile characterisation—Implications for demand side management. *Energy* **2019**, *180*, 665–677. [CrossRef]

46. Hajiaghapour-Moghimi, M.; Azimi-Hosseini, K.; Hajipour, E.; Vakilian, M. Residential Load Clustering Contribution to Accurate Distribution Transformer Sizing. In Proceedings of the IEEE 2019 International Power System Conference (PSC), Tehran, Iran, 9–11 December 2019; pp. 313–319.
47. Azaza, M.; Wallin, F. Smart meter data clustering using consumption indicators: Responsibility factor and consumption variability. *Energy Procedia* **2017**, *142*, 2236–2242. [CrossRef]
48. Grigoras, G.; Scarlatache, F. Processing of smart meters data for peak load estimation of consumers. In Proceedings of the 2015 9th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, 7–9 May 2015; pp. 864–867.
49. Schubert, E.; Rousseeuw, P.J. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. In *Transactions on Petri Nets and Other Models of Concurrency XV*; Springer Science and Business Media LLC: Cham, Switzerland, 2019; Volume 11807, pp. 171–187.
50. Zainab, A.; Syed, D.; Ghrayeb, A.; Abu-Rub, H.; Refaat, S.S.; Houchati, M.; Bouhali, O.; Lopez, S.B. A Multiprocessing-Based Sensitivity Analysis of Machine Learning Algorithms for Load Forecasting of Electric Power Distribution System. *IEEE Access* **2021**, *9*, 31684–31694. [CrossRef]
51. Villar, J.; Bessa, R.; Matos, M. Flexibility products and markets: Literature review. *Electr. Power Syst. Res.* **2018**, *154*, 329–340. [CrossRef]