Electronics and Electrical Engineering

http://www.aimspress.com/journal/ElectrEng

AIMS Electronics and Electrical Engineering, 5(4): 284–314.
DOI: 10.3934/electreng.2021015
Received: 01 September 2021
Accepted: 18 November 2021
Published: 26 November 2021

Review

A systematic review of data pre-processing methods and unsupervised mining methods used in profiling smart meter data

Folasade M. Dahunsi[1,]*, Abayomi E. Olawumi[2], Daniel T. Ale[2] and Oluwafemi A. Sarumi[3]

[1] Department of Computer Engineering, The Federal University of Technology, Akure, PMB 708, Akure, Ondo State, Nigeria
[2] Department of Electrical and Electronics Engineering, The Federal University of Technology, Akure, PMB 708, Akure, Ondo State, Nigeria
[3] Department of Computer Science, The Federal University of Technology, Akure, PMB 708, Akure, Ondo State, Nigeria

* **Correspondence:** Email: fmdahunsi@futa.edu.ng.

Abstract: The evolution of smart meters has led to the generation of high-resolution time-series data - a stream of data capable of unveiling valuable knowledge from consumption behaviours for different applications. The ability to extract hidden knowledge from such massive amounts of data requires that it be analysed intelligently. Hence, for a clear representation of the various consumption behaviours of consumers, a good number of data mining technologies are usually employed. This paper presents a systematic review of the various data mining techniques and methodologies employed while profiling energy data streams. The review identifies the strengths and shortcomings of existing data mining methods as applied in research, focusing more on data processing techniques and load clustering. Also discussed are data mining methods used to profile consumption data, their pros and cons. It was inferred during the research that the choice of data mining technique employed is highly dependent on the application it is intended for and the intrinsic nature of the dataset.

Keywords: load profiling; data mining; energy data; smart meters; data pre-processing; load clustering

1. Introduction

With the growing interest in Smart Grid, consumption data in electricity collected are no longer limited to only billing and pricing purposes [1,2]. The need to measure electricity consumption data coupled with the recent advances in sensing, measuring and control technologies has also led to smart meters that can measure and communicate electricity consumption [3,4]. Smart meters are capable of collecting energy data parameters at a higher frequency (15 minutes or less), providing a large volume of load data that is capable of giving a quality reflection of consumers' behaviours compared to traditional meters [5–7]. The data on electricity consumption collected are forwarded through reliable communication systems to central collection points, usually via a Wide Area Network (WAN) [8,9]. Analytics are then carried out on the data to derive inherent valuable knowledge that can be used to make informed decisions that will improve the efficiency of the energy systems, such as implementing some demand response programs [10]. As illustrated in Figure 1, researchers carry out data pre-processing, cleaning, and profiling to derive rich and related knowledge from the data on electricity consumption. In some scenarios, analysts augment the consumption data with other sociodemographic data features such as weather information. Also, in order to simplify the data for analysis, features are often extracted from the data before analysis [11].

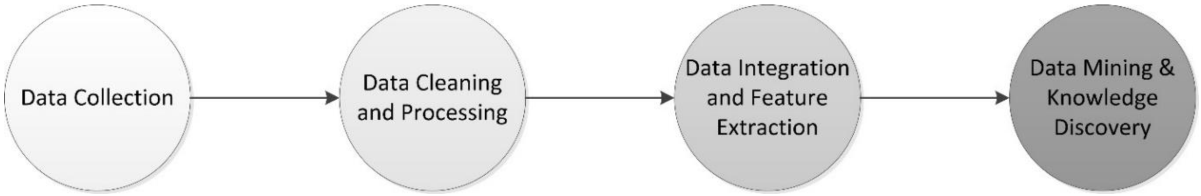
Data Collection
Data Cleaning and Processing
Data Integration and Feature Extraction
Data Mining & Knowledge Discovery

Figure 1. Knowledge discovery process.

Data analysis for profiling and classifying customers based on their energy usage pattern is important for demand and supply-side management. Load Profiling (LP) involves the classification of load curves according to consumption patterns over a period [11,12]. It is an essential tool for regulators in their bid to manage energy systems. Accurate load profiling leads to better load scheduling, peak-load detection and other demand-side management applications [13–16]. However, mining energy data has become exerting and expensive, following the phenomenal growth in energy databases and the stochastic nature of loads [17,18], such that it is becoming impossible to analyse consumer consumption behaviour at an individual level [19,20]. Hence, [6], [21], and [22] noted that traditional databases and statistical analysis are no longer sufficient in extracting knowledge from raw data coming from smart meters. The conventional approach involves aggregating the time series data into a representative profile using simple statistical analysis, but this is impractical for profiling or grouping consumers [23]. This approach lumps different consumption behaviours together, hence giving poor and misleading Load Profiles (LPs), which in many cases leads to the loss of relevant information about the consumers.

Furthermore, wireless meters occasionally lose data packets due to power loss, transmission error or unstable internet connection [4,9,24–26]. However, most challenges posed by big data occur during data preparation [26]. Therefore, data analysts must adequately address missing values, outliers, duplications in the consumption data received and other related issues.

Data Mining (DM) techniques are the most auspicious tool for deriving knowledge from large datasets and are classified as: supervised and unsupervised analytics [14,27,28]. DM techniques are very useful in technically addressing the highlighted complexities without altering the quality of extracted embedded information [29]. Hence, DM techniques such as clustering are instrumental to load profiling analysis. It involves intelligent grouping of consumers based on their consumption behaviours, identifying similar electricity consumption patterns before any aggregations are applied. This way, there is no loss of profile shaping [22,23]. After clustering, the cluster's profiles are aggregated by evaluating the shape characteristics of each cluster or by aggregating the centroids from each cluster depending on the objective(s) of the clustering, thus, giving representative LPs [30]. Furthermore, the clustering of various consumption behaviours (clusters) of a consumer (knowledge discovery) is carried out to obtain LPs for a single household.

Most previous studies investigated the various techniques used in collecting and applying smart meter data. Only a few paid close attention to the techniques employed during data preparation and load profiling. A recent comprehensive review on smart meter data analytics was conducted by [11]. However, it only reviewed literature around applications, methodologies, and challenges of analysing consumption data. Also, [31] reviewed load profiling and its application to demand response programs such as price-based and incentive-based programs. The paper excluded issues around data preparation. This paper differs from [11] and [31] by focusing on the data mining methodologies employed during data preparation and clustering of smart meter data towards load profiling. Association rule mining, which is potentially valuable for information extraction and knowledge mining from big data, was not considered because its application to smart meter data is relatively new. Only a few articles employed the mining technique [32]. This paper emphasises data pre-processing and load profiling using unsupervised mining methods to fill the gap. It provides an updated review of data mining techniques employed for this purpose. It analyses and evaluates existing and novel approaches adopted on outlier detection, data normalisation, data reduction and load profiling.

Pre-processing of data is a fundamental task that transforms acquired raw data into more useful and understandable format. Inefficient data pre-processing leads to poor and misleading load profiles and data interpretation [25,26]. Adequately addressing issues such as data missing values, outlier detection, data reduction, data normalisation, data duplication, and other related issues leads to effective detection of information from data that are relevant and useful for stakeholders to make informed decisions.

Section 2 of this paper illustrates the methodology employed in the acquisition of related articles that were reviewed. Section 3 presents the results obtained from the systematic review conducted. Section 4 discusses the knowledge obtained from the reviewed articles, which includes the various data mining techniques employed during data pre-processing and data clustering towards the profiling of load consumption data, while Section 5 presents the conclusion that can be drawn from the review.

2. Methods

This paper conducted a systematic review following the reporting checklist of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [33]. A bibliometric analysis was carried out on the IEEE Xplore and ScienceDirect databases on 12 January 2021 and 28 January 2021. The query employed for both databases is as follow: ((“smart meter” OR “consumption”) AND “data” AND “mining” AND (“technologies” OR “methods”) AND “load” AND “profil*”). The query was used to explore the ‘article title, abstract, keyword, the content’ of every published document in the database between 2010-2021, following the fact that the subject area is a novel and growing field and the smart grid initiatives started around the late 2000s [11]. In addition, this review only considered research articles around data mining and smart meter data profiling.

3. Results

The first search from the IEEE Xplore and Science Direct databases gave 2,907 and 23,080 papers, respectively. The authors screened the papers carefully to 99.

The search result from the IEEE Xplore was first indexed to extract only conference papers, journals, early access articles and books, reducing the articles to 2884. During the selection phase from the two search results, the paper excluded some "consumption data" related articles such as wind power plants and photovoltaic power systems as they are outside the scope of the review; this also reduced the articles from the IEEE Xplore database to 1810. The reviewers filtered out 1736 after reading through the title and abstract, leaving 61 related articles from the IEEE Xplore database. On evaluating the results from ScienceDirect, considering only review articles, research articles and book reviews and reading through the titles, 31 related articles were selected.

The review considered articles that combined other socio-economic and demographic factors with acquired smart meter data when profiling electricity consumers on both databases. These have a significant effect on consumers' profiling results. Also excluded are articles with in-depth consumption profiling such as household appliances identification and profiling as such subjects are broader than the scope of this review. In the sorting, google scholar was used to furthering select 11 other articles found pivotal to understanding some of the articles obtained from the two significant databases, yielding 121 reviewed articles.

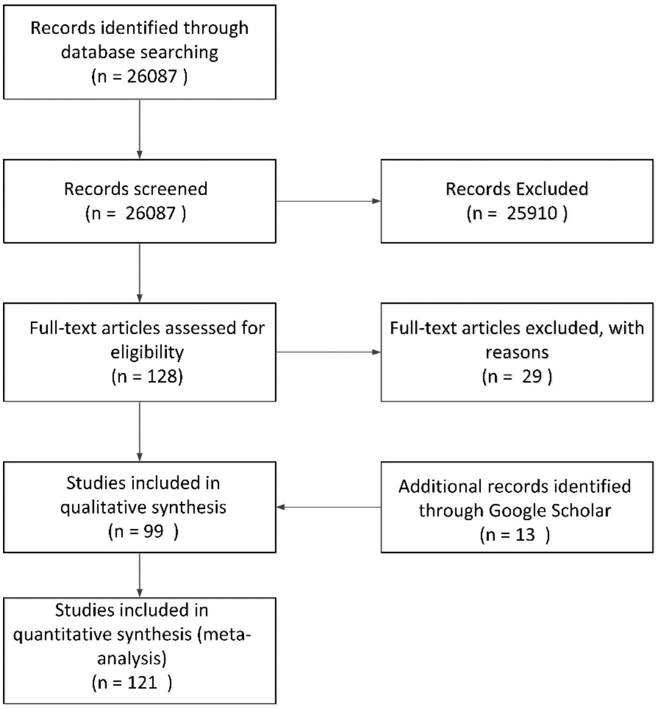
Records identified through database searching (n = 26087)
Records screened (n = 26087)
Records Excluded (n = 25910)
Full-text articles assessed for eligibility (n = 128)
Full-text articles excluded, with reasons (n = 29)
Studies included in qualitative synthesis (n = 99)
Additional records identified through Google Scholar (n = 13)
Studies included in quantitative synthesis (meta-analysis) (n = 121)

Figure 2. Flow chart of the study selection process.

4. Discussion

4.1. Data pre-processing

The quality of high-resolution consumption data does not infer its direct usability. Data

Pre-processing is the process of preparing data for knowledge discovery. It consists of tasks such as cleaning, transformation, integration and reduction due to incompleteness, noise, ambiguity and inconsistency that are obtainable in a dataset [10,34,35]. Data collected must be cleaned, transformed and(or) reduced before they can be easily interpreted, visualized and manipulated [36]. Real-world data are generally highly susceptible to inconsistencies; hence the need to clean consumption data from issues such as missing data and outliers remains fundamental [8,29]. Also, for smooth analysis, every column in a dataset is expected to take a standard data format (such as timestamp: date format, energy consumed: a processable format). Many technical works of literature are silent about the techniques employed in data cleaning [37], while some works of literature such as [38] decided to use uncleaned data, noting that cleaning the data obtained may influence the estimation process. A substantial percentage of works of literature were silent about their approaches to missing data as some made use of already pre-processed datasets [18,39]. However, for practical application of any knowledge embedded in raw smart meter data, the data must be cleaned from every abnormality as careful pre-processing can improve the result of data analysis [40]. Discussed in this section are the major setbacks and the various techniques employed in reviewed literature.

4.1.1. Missing data treatment

Missing data is a common issue when processing a dataset. Missing data often arises from faults in sensors, technical issues, instability in internet connection or malfunctions from any electronic components in a smart meter [35]. Any decision taken on missing data comes with consequences, as missing data impacts learning, inference and prediction obtainable from the data, depending on its proportion [25]. Therefore, before adopting any approach, it is essential to determine the nature and pattern of the missing data (random or non-random) in the complete set. This assists in selecting the appropriate approach to adopt in handling missing data. Figure 3 illustrates the various approaches to missing data. The missing values in a dataset can either be ignored or filled. Deleting or ignoring data is the simplest solution, but it is not advisable,it may contain valuable information about consumption behaviours. This approach is appropriate when the missing data is relatively small compared to the total population, considering that the process might drop a single value or entire row(s) [41]. Z.A. Khan et al. [18] also noted that it might be better to remove missing data rather than fill them using exploration techniques. Hence, many works of literature neglected a subset of the total population of their dataset due to missing data to avoid uncertainties and preserve the original features of data [20].

Replacing missing can be done through general methods such as imputing an evaluated value or using the advanced approach, which involves subjecting the data (some portion) to models such as K- Nearest Neighbour (KNN). A common approach for non-time series data is mean imputation. It involves the replacement of the missing value with the averaged value. This method is strongly criticised. It is known to have a significant influence on the data’s variance [42]. In replacing a time series data which the nature of consumption data, the backward fill and the forward fill are very ubiquitous, the “backward fill” approach involves the replacement of the missing value with the previous reading, while the “forward fill” approach propagates the missing fields with the last valid values forward. It is important to note that the techniques mentioned above depend on themselves, not other neighbouring variables, according to [43]. This approach often leads to inefficient analysis and produces biased estimates of the association investigated. However, techniques such as interpolation and regression allow for the maximisation of other dependent and independent variables during data imputation. Illustrated in Table 1 is the nature of the consumption data experimented on by some works of literature and the techniques employed in handling missing data.

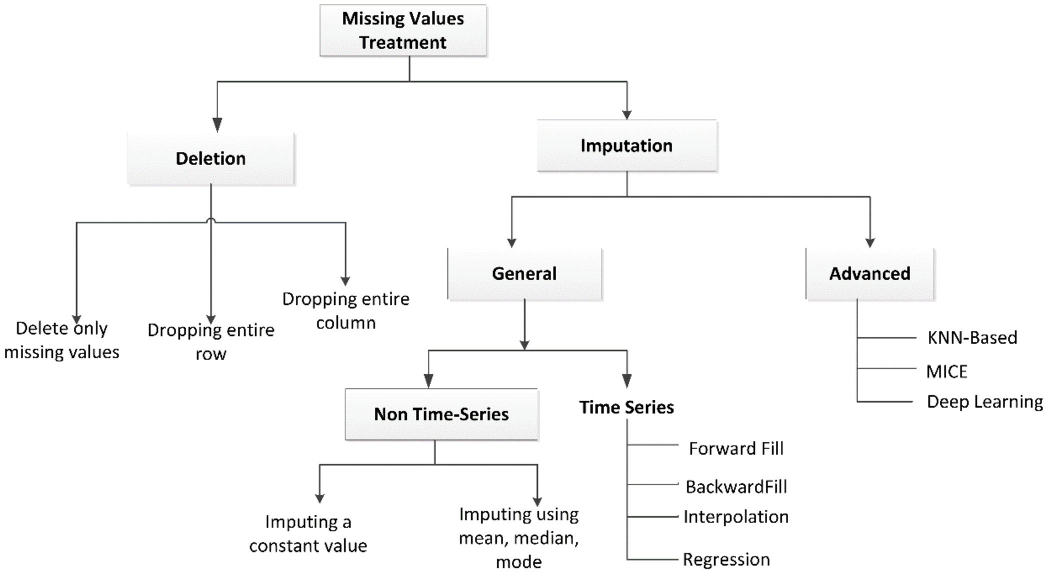

Figure 3. Dendrogram of general approaches to missing data.

Table 1. Approaches of literature to missing data.

Literature	Period of Collection/Total Size	Missing Data (%)	Method employed
[44]	1 year / 660 users	not stated	Linear Interpolation
[45]	not stated / 208 users	not stated	Multilayer perceptron (MLP) Artificial Neural Network
[41]	1 year/ not stated	not stated	deletion
[46]	1 year / 4554 customers	not stated	deletion
[47]	537 days / 6445 users	0.18%	deletion
[18]	not stated / 5000 users	0.94%	deletion
[22]	6 months / 4000 users	1.25%	deletion
[48]	1 year / 15433 users	not stated	deletion
[35]	not stated / not stated	1.76%	Improved Lagrange Interpolation (PB’Eyes)
[49]	900 days / 100 users	13.85%	deletion
[40]	7 days / 34418 users	0.2%	deletion
[20]	366 days / 200 users	17.5%	deletion
[29][50]	1 years / 171 users	not stated	Single exponential smoothing & SARMA model
[30]	1 year / 105 buildings	23%	deletion
[42]	1 year / 30 users	not stated	Mean Imputation & deletion
[51]	1 year / 106 users	17%	deletion
[52]	1 year / 269 users	15%	Inference Method

Among the general methods illustrated in Figure 3, interpolation is the most used [35,44]. It involves deriving unknown values from a known set of values in the dataset. Among the various mathematical formulae for interpolating, the Lagrange interpolation was well adopted [35]. Another ubiquitous approach in Table 1 is the replacement of missing values with the average load of a highly correlated time interval.

Another form of compromise was found in [53], where attention was only given to a series of missing data with not more than five (5) days in a row using the inference method. Missing data were replaced with data points of the previous week (same day, same hour)[52,54], neglecting the rest. J.P. Gouveia et al. [42] used a dual approach: deleting from the total population, i.e. households whose missing data is more than 20%, and subjecting other cases to mean imputation methods (using neighbouring values). A. Mutanen et al. [44] employed linear interpolation to fill missing data when the data interval is not more than five (5) rows, while any pattern of missing data more than five (5) rows was removed from the set.

A good number of general methods have been modified to suit individual researchers' purposes; an improved linear interpolation model called moving window was found in [54]. They noted that the method is easy to implement when the duration of the missing data is small. P. Manembu [25] also developed an improved Lagrange interpolation called PB'Eyes. This solution was used to obtain the data pattern of a missing set by subjecting any missing row in the dataset to either first order, second order or third order Lagrange Polynomial based on the numerical relationships between the missing row and the surrounding rows.

Besides the conventional methods illustrated in Table 1, many recent works of literature have developed unique and novel approaches to handling missing consumption data. An approach called forecasting, and exponential smoothing technique was found in [29] and [50]. This method employed the single exponential smoothing equation when the missing value window in the dataset was more than 12, and Seasonal Auto-Regressive Moving Average (SARMA) model was utilised when less than 12. However, a recent comparison was made by [55] using power data which still infers the superiority of machine learning models over statistical methods. The authors compared two statistical methods: autoregressive integrated moving average (ARIMA) and linear interpolation (LI) models, and three machine learning methods, KNN, multilayer perceptron (MLP), and support vector regression (SVR). More novel and sophisticated approaches to replacing missing data were found in [43,56–60]. In a recent work by [43], it was noted that the interpolation technique requires relevancy between the neighbouring variables and the variable of interest, which may not be obtainable in all cases. Also, the regression technique requires that the form of function be pre-defined.

Consequently, the parameters of the function need to be estimated by model training. Hence, the authors considered the interpolation and regression methods inefficient as they could not capture the trend. This conventional setback led to the adoption of a non-linear compensation algorithm for a linear strategy of replacing missing data [43,58–60]. This method was proved by [43] and [60] on energy data, but it has not been well adopted by various works of literature related to smart meter data.

4.1.2. Outlier detection

Outliers are data points whose values are exceptionally far from the mainstream data value; this often results from measurement error or data corruption during communication [10]. Such data points could take a negative value or an overly-high value [61]. This data set gives a wrong definition of consumers’ behaviours if not technically identified and managed. Due to the enormity of consumption data, identifying outliers manually by visual observation can be exerting and misleading; hence, the need for data mining techniques. Having identified the outliers from a set, they could be dumped or replaced using any of the approaches discussed in the previous section. It was observed that not all outliers are errors. Some outliers bear some pieces of information about consumers’ behaviours or the power grid, from which relevant knowledge can also be drawn [38,62]. During the pre-processing of data by [44], the authors opined that having outliers in load data does not necessarily imply the measurements were erroneous but noted the outliers might not be suitable for some applications such as anomaly detection [63]. Having identified outliers in data, they can be removed or replaced using any treatment techniques discussed in the previous section. Authors in [64] replaced some outliers using linear interpolation. They identified and removed the lower and upper outliers of the consumption distribution using the Interquartile Range (IQR) of the standard deviation and K-Nearest Neighbour (KNN), respectively.

In [7], they employed three (3) algorithms named distance-based, density-based, and Local Outlier Factor (LOF) to identify outliers in the REDD dataset used for consumption analysis. The three (3) algorithms gave a similar result, which filtered out 10 out of the 443 houses considered in the research. The authors justified the removal because inclusion will influence the clustering analysis further done on the entire data set. In [65], daily consumption data points that fall outside three standard deviations (3σ) were considered outliers. X. Lin et al. [66] assumed the load data to be a normal distribution and set the reject level at $\alpha = 0.05$. A. Mutanen et al. [44] also followed this assumption, customers whose monthly consumptions differ from the average consumption of all customers, outside the given probability (80% - 99.99%) from the normal distribution were regarded as outliers and therefore not profiled. Furthermore, after obtaining the daily load profile in [41] from averaged load diagrams for each consumer, the authors noticed that some days were outrageously different from the representative load diagram. The authors dealt with outliers by discarding 10% of days with the highest Euclidean error [44].

Evaluating a set’s interquartile range (IQR) is another way to fish out outliers [67]. However, J. Yang et al. [68] opined that instead of marking points that lie alone in low-density regions as outliers. Furthermore, some outliers are never evident until the data are clustered, while some are identified while formulating the representative profile for each consumer [69]. Hence, in recent research works such as [62] and [70], the total smart data accrued were first subjected to a clustering technique to remove outliers. Z.A. Khan et al. [71] filtered out 0.31% outliers after subjecting the pre-processed data to extended K-means clustering. After subjecting the data to K-Means clustering [50] removed outliers by computing an upper-lower band for every consumer cluster, using three scaled median absolute deviations to calculate the upper and lower bound using minimum sample points of all load profiles in a cluster. Load profiles outside this bound were considered outliers in the formation of the representative profile for each cluster. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is the most recent clustering technique found in works of literature for detecting outliers. The algorithm classifies data points with insufficient neighbours as noise. X. Liu et al. [23] validated the efficiency of the hypothesis by using the DBSCAN algorithm to detect outliers of daily electricity load profiles before subjecting it to a second clustering step for Typical Load Profile (TLP) extraction. D.I. Jurj et al. [62] also carried out a comparative analysis of Interquartile Range/ Median (IQR), LOF and DBSCAN methods. The results showed DBSCAN algorithm as a superior technique for outliers’ identification among many other techniques.

4.1.3. Data normalisation

With the diverse and complex consumption data, ML algorithms will not function correctly without scaling down and confining the original consumption data to a specific range. Thus, for practical analysis of consumers' behaviours, it is imperative to set the total population within a standard; with this in place, all data points are ably represented in the complete set [20,36,53,72]. Normalisation aims to reduce the magnitudes of the data sequence without altering the consumption behaviours. [71] and [73] noted that the normalisation's choice affects the clustering result as some methods fail to maintain the original definition after scaling down. Hence, the method adopted in any application work should be wisely selected. Generally, there are three standard methods of normalisation: decimal scaling, min-max normalisation, and z-score normalisation. However, the two latter methods are the most used in energy data analytics [20,74].

Minimum-Maximum or Unity based Normalisation: This method ensures all data are linearly mapped into the interval [0] – [1]. Unity-based normalization scales consumption data down to the range [0-1] before clustering, however, different equations as illustrated in Table 2 were seen in literatures but Equation (2) remains the conventional expression, used in [13][20][29][38][40][41][47][66][71][75][76][77]. During the hourly load profiling carried out on 112 feeders by [10], power data were normalized to the range 0 -1 using Equation (4); while ensuring that equation (5) is maintained. Furthermore, [23][36][41][51][64][77][78] and [79] normalised the readings from each customer in their works by dividing every data point from each consumer by their respective maximum power value (which could be daily, monthly or yearly load consumption), [41] noted that the method enables the comparison of the consumption profile with other customers, regardless of the consumption volume of each one.

Z-score Normalisation: This normalisation method scales the data sequence using the mean and standard deviation of the data sequence. From our review, only a few works of literature employed this approach due to its high sensitivity to outliers [65]. However, [80] noted that z-score normalisation has no adverse effect on analysis where usage pattern study is insignificant. Equation (7) in Table 2 presents the expression for this approach. This method is not as effective as Unity-based Normalisation. [30] compared the impact of z-score and min-max normalisation on daily and annual profiles, the result showed that many operational intensities were lost using z-score normalisation.

Other approaches to normalisation akin to min-max normalisation were found in some works of literature. Different corresponding values were employed rather than normalising each data point by dividing by the respective maximum data point. [53] normalised its daily profile using the average of the data point of each day. [30] presented a varied approach during the daily load profile formulation over one year using a sub-method called maximum normalisation. The hourly raw consumption was normalised by dividing each data point by the maximum value in the year instead of the day. The authors noted that the method is more preservative as it helps maintain information about the base load. Also, a novel method called 'de-mining was found in [84]. The minimum load demand was first subtracted from each data point before dividing each by the total consumption, following the fact established by [85] regarding the 'de-mining method's high efficiency when linking exogenous activities to electricity consumption patterns for analysis.

Table 2. Normalisation techniques.

Normalization Techniques	Literatures	Equation	Variable Definitions
Unity or Min-Max Normalization	[20][29][38] [40] [41][48] [51][61][71] [66] [75] [76] [79]	$x^i = \frac{x - min(x)}{max(x) - min(x)}$ (2)	x =original load data sequence, min(x) = maximum values of the load data sequence, max(x) = maximum values of the load data sequence, x' = normalized data sequence.
	[13]	$Y_t = A_w . P_t + e_t$ (4) $\sum_{t=1}^{24} P_t = 1$ (5)	Y_t = Power drawn by a feeder at a time A_w = Average Power of the day e_t = Error in power drawn by the feeder P_t =Proportion of Daily Load that is drawn the tth hour interval
	[9] [23] [30] [36][41] [51][64] [78] [79] [81] [82]	$N_{L_d,t} = \frac{L_{d,t}}{L_{max}(year\ or\ month\ or\ day)}$ (6)	$N_{Ld,t}$ = the normalised load at a specific hour, day d and time t, $L_{d,t}$ = the original load at the corresponding hour, day and time, and, $L_{max\ ,year}$ = the maximum load data during the year or month or day at the corresponding day and time
Z- Score Normalization	[65][80][83]	$v^i = \frac{v - u}{\sigma}$ (7)	v =original load data sequence, μ= mean value load data sequence, σ= standard deviation of the load data sequence, v' = normalized data sequence.

4.2. Data reduction techniques

Reducing consumption data before analysis has the tendency of reducing the cost of computing of high-resolution consumption data and also improving clustering computing performance [19][49][65][81], hence, the existence of numerous compression techniques. Compression techniques generally have been classified into two (2) major classes: the lossy and the lossless compression techniques [86–88]. Lossless techniques are known to produce a quality compression ratio at the expense of information loss. While the lossy techniques maintain the reverse order as they prioritise reducing the information loss at the expense of the compression ratio. Among the various lossy techniques are the Discrete Wavelet Transform (DWT), the Discrete Fourier Transform (DFT), Piecewise Aggregate Approximation (PAA), Principal Component Analysis (PCA), Singular Value Decomposition (SVD). However, the lossy compression techniques are the most adopted as they are more useful in accelerating a similarity search and extracting useful information [20]. Hence, they are mostly used in dimension reduction before clustering [31].

According to [89], consumption data reduction can either be raw-data based, feature-based or model-based. However, the feature-based and model-based remain the mostly adopted [9][23][29][53][61][80][90][91].

According to [29], the feature-based method is less sensitive to missing data and can effectively reduce raw data having different resolutions with minimal loss [29][73][92]. [29] referred to it as a pre-processing step needed for improving clustering, while such is generally referred to as Indirect clustering. Feature extraction involves extracting features such as mean, standard deviation and some load profile characteristics such as load factor from the raw data without changing them, still giving the true definition of the raw data. They can be extracted from statistical computation or via a learning algorithm.

The extracted features are then further normalised and used for clustering [61]. Using feature extraction, [31] classified indirect clustering methods into dimension reduction-based clustering and time series-based clustering. The former uses linear dimension reduction methods such as PCA and deep learning to replace consumption data with artificial variables. The latter uses analytical methods to extract features from time-series data, employing PAA, SAX, SVD etc. Extracting the frequency domain representation (consisting of the amplitudes and phases) of the time series data using algorithms such as Fast Fourier Transform (FFT) is another approach to feature extraction found in the literature [93,94]. However, despite the effectiveness of the feature-based approach, the residential consumption load dataset is bound to contain noise and unequal time series, a challenge that has a significant influence on clustering results [94].

In addition to the need to address issues around privacy, some researchers prefer the model-based approach. The effectiveness of the model-based approach was validated by [95]. The approach was subjected to an incomplete and noisy time series; the authors noted that the approach tends to facilitate parallel computing and fast clustering. The result obtained by [29] using feature extraction before clustering was compared to that from Self-Organizing Mapping (SOM) based clustering using the following performance indices: Davies–Bouldin Index (D.B. index), Calinski–Harabasz index (C.H. index) and Silhouette score (S score). The result shows that feature extraction as a means to data reduction is far more efficient than direct clustering. However, the method is application dependent. [53] also validated the superiority of feature-based clustering by comparing the clustering performance obtained when five (5) features were extracted before clustering and the direct clustering using the Silhouette score (S score).

A. Tureczek et al. [40] noted that some principal components such as intrinsic temporal data structure hidden in the smart meter data are often overlooked by clustering algorithms (in their case, K-Means) if such prominent features are not extracted before clustering. Hence, they employed normalisation, wavelet transformation and autocorrelation in their work to extract the temporal components before clustering. The result indicated a better clustering and faster performance than direct clustering.

Y. Wang et al. [39] decomposed daily load profiles into a few partial usage patterns (PUPs). They used sparse coding and non-negative K-SVD technique to compress consumption data and extract PUPs, respectively, due to the sparse nature of the dataset. The authors noted that high electricity consumption only occurs relatively at a small fraction of time for residential consumers, leaving consumption for the remaining period as zero. The PUPs extracted were finally combined linearly to form a consumption load profile. The K-SVD achieved a better compression ratio and lower information loss when compared with other lossy compression techniques (PCA, PAA, DWT). However, the time required for coding the K-SVD was 31 minutes higher than PCA and DWT. Hence, they opined that the technique needs optimisation for real-time applications.

S. Lu et al. [20] and Y. Shi et al. [65] made use of Piecewise Aggregate Approximation (PAA) to transform dense uncompressed load data into piecewise paned data. PAA performs reduction on time series data by obtaining the mean value of each subsequence, which thus becomes the representation of the sequence. Using PAA, the number of segments, the number of data points represent each time series, is provided before evaluation [65]. Also, in [20], the weekly load data consisting of 15-minute (i.e., 672 dimensions) were compressed into 21 dimensions using PAA. This was achieved by dividing each day's load data into three representative data values, covering the average load (mean value) of 0:00-8:00, 8:00-16:00, and 16:00-24:00 daily. The accumulated evaluation from each day was then approximated as the week load profile. The efficiency of PAA was not compared with other compression techniques. However, the weekly load profile was clustered and used to create a model for characterising each cluster. In [65], the raw consumption data of each consumer was reduced by setting the number of segments to eight (8), reducing the raw daily load data. Symbolic aggregate approximation (SAX) was then introduced to take care of possible abnormalities in the obtained profile that may have resulted from faults or reading errors from the meter; thus, reducing the data to categorical data. Illustrated in Table 3 are the various reduction techniques employed by considered works of literature.

Table 3. Data reduction techniques.

SN	Data Reduction Technique	Literature
1	Discrete Fourier Transform (DFT),	[37]
2	Symbolic aggregate approximation (SAX)	[47][81]
3	Principal Component Analysis (PCA)	[79][96]
4	Singular Value Decomposition (SVD)	[39][97][98]
5	Piecewise Aggregate Approximation (PAA)	[20][65]
6	Self-Organizing Map (SOM)	[89]
7	Convolutional Autoencoder (CAE)	[80]
8	Fast Fourier Transform (FFT)	[93][94]
9	Variational Recurrent Autoencoder enhanced with LSTM neural networks	[87]
10	t-SNE (TDistributedStochastic Neighbor Embedding)	[30]
11	Multidimensional Scaling (MDS)	[69]

4.2.1. Cons of data reduction techniques and way forward

According to [18][19][39][86] and [99], under certain circumstances, reducing the dimensionality of an input dataset affects the clustering result. It alters the accurate representation of the original raw data as some pieces of information are lost during reduction. Also, [19] highlighted that the accuracy of any reduction technique is highly dependent on the intrinsic nature of the raw data, such as the number of readings captured in each time interval in the time-series data. Therefore, different lossy compression techniques have different impacts on different applications. The compression technique to be selected should be carefully chosen according to the characteristic nature of the dataset (i.e., its sparsity and diversity) [39]. However, novel approaches are evolving to reduce the size of time series data with minimal loss on the information.

It was observed that feature extraction done by the majority of works are manual [90]. Considering the variability and non-linearity of individual load data, this is highly ineffective; thus, [80][91][87] and [100] recommended the use of deep learning techniques for feature extraction. The efficiency of this technique was experimented with in [91] when they used Convolutional Neural Network (CNN) to extract features from energy data before subjecting it to Support Vector Machine (SVM). They were identifying the socio-economic characteristics of consumers. The result obtained gave the best accuracy when compared with seven (7) other methods: Biased Guess (BG), Manual Feature Selection (MF), SVM, L1-Based Feature Selection + SVM (LS), PCA + SVM (PS), Sparse Coding + SVM (SS), CNN + Softmax (CS). [80] also employed a deep learning-based framework (Convolutional Autoencoder, CAE; a combination of an autoencoder and CNN) to reduce a Yearly Load Profile (YLP) of 8460-dimensional space to 100-dimensional vectors. The approach gave a high compression ratio of 130% and less reconstruction error when compared with other dimensionality reduction techniques (PCA, kernel PCA, Independent Component Analysis (ICA), autoencoder (AE) and DWT).

4.3. Clustering

Clustering is one of the popular techniques employed while profiling streams of load data. It involves identifying and grouping consumers with similar consumption behaviour into classes, with each class aggregated into a single profile, often called Typical Load Profile (TLP) [37][101][102][103], which then becomes the representative profile of the consumers in such class. Also, as earlier discussed in the previous section, clustering load data is very instrumental in identifying outliers in a dataset [46][90]. Through works of literature, consumption data take any of the three (3) different forms before clustering can be performed [29][104]:

1. Raw time series consumption data
2. A reduced dataset was obtained from the raw dataset using DM techniques
3. Features extracted from consumption data

Contained in this section are the justifications for any of the listed forms. Generally, Clustering is classified into direct clustering and indirect clustering [31]. Direct clustering involves the direct application of a clustering technique to raw data. On the other hand, indirect clustering involves clustering features extracted from raw or reduced data [18][31]. Direct clustering can be time-consuming, and high dimensional data can result in anoverfitting of the clustering algorithm. Hence, many works reduce the dimension of the pre-processed data using some dimension reduction techniques before clustering [18]. However, [89] noted that clustering using raw data comes with a clear advantage in performance; hence, in their work, they selected 5% of the original data for clustering. [29] noted that dealing with the high-dimensional consumption of raw data by down sampling or aggregation often leads to information loss. A good number of recent works of literature have adopted clustering techniques in load profiling. However, disparities were observed in the approaches such as (i) period of study, (ii) clustering algorithms adopted, (iii) data reduction techniques (iv) the use of features and many more [104]. Discussed in this section are the variants and justifications. However, due to the uneconomical cost on computational resources, while clustering too many load, optimisation algorithms such as seen in [102] are developed to mitigate the challenge.

4.3.1. Clustering algorithms

Numerous data mining techniques can be used to group consumers into classes based on consumption behaviours. However, some recent research works tend to employ more than one type of clustering algorithm this approach is called second-order clustering. The approach tends to provide better Typical Load Profiles (TLPs) in some works, depending on the application in view [23][30][51][105]. In the case of [66], the authors employed hierarchical and fuzzy c-means and noted that the fuzzy c-means algorithm is susceptible to initial clustering centre selection. Thus, they classified the load characteristics using the hierarchical method before subjecting it to fuzzy c-means. By so doing, the clustering centre was provided by the hierarchical clustering, which consequently influenced the accuracy of the result obtained. [89] also employed three (3) clustering algorithms (SOM, K-Means and Hierarchical) while generating load profiles from many customers. The approach was adopted based on the fact that; clustering using SOM reduces the size of the data, and it is computationally effective than direct clustering. The result obtained indicated that SOM + K-Means has better performance than SOM + Hierarchical clustering.

Most clustering analyses run on real power consumption, not considering the impact of power quality. A novel approach was found in [106], where the K-Means clustering algorithm and Fuzzy logic were used to cluster real power consumption and Total Harmonic Distortion (THD) pattern separately and later combined. Generally, a conventional clustering algorithm cannot handle mixed data. Hence, Fuzzy Logic Clustering: an algorithm that can handle multiple patterns, was introduced. The result obtained by the authors gave a more precise reflection of the consumption behaviours of consumers.

Multiple-step clustering was also found in works of literature. Step clustering involves the subjection of a dataset to a clustering more than once, using the same algorithm. This approach can be adopted for outliers removal before subjecting the processed set to another level of clustering for TLP identification, this type of clustering is called second step clustering – clustering twice [30][69]. It is mainly adopted for a deeper level of grouping. In [69], intra-building clustering was carried out on every building in a complete set using Gaussian Mixture Model (GMM) for outlier removal. After that, inter-building Hierarchical clustering for extraction of TLP of all buildings was carried out. The introduction of this strategy with a reduction technique reduced the computational cost by the ratio of 16.5:1 compared to the direct hierarchical clustering-based strategy.

This subsection discusses the working principles, strengths and weaknesses of some of the standard clustering techniques obtained in works of literature.

a) K-means Clustering

Partitional clustering algorithms are known for quality output due to the iterative optimisation process employed by them. Among the various partitional clustering algorithms, K-Means clustering is the most employed partitional clustering algorithm applied to load-profiling [15][19][45]. It is highly scalable and can sort large datasets relatively fast [106–108]. The clustering process for K-means starts by first carefully selecting the number of clusters, K and the initial centroids. Each data point in the remaining population is assigned to the closest centroid depending on the adopted proximity measure. On the formation of the clusters, the centroid of each cluster is recomputed. The last two (2) procedures are repeated continuously until the centroid stops changing [14][73]. The need to pre-set the number of clusters before clustering is always a challenge. Hence, the adoption of automatic assignment of the number of clusters and the possibility of validating a clustering output to improve clustering performance led to the modification of K-Means [14]. As illustrated in Figure 4, researchers lately adopted the idea of integrating the validation check as part of the clustering process [27]. The modification of the K-Means algorithm using this approach is referred to as adaptive-K-means. [15][20] used the adaptive K-Means, the evaluation of DBI was integrated as part of the process. The lesser the value of the DBI, the better the clustering result. The number of K were continuously and automatically increased until the DBI of the clusters no longer decreased. In this way, the best number of clusters, K was selected.

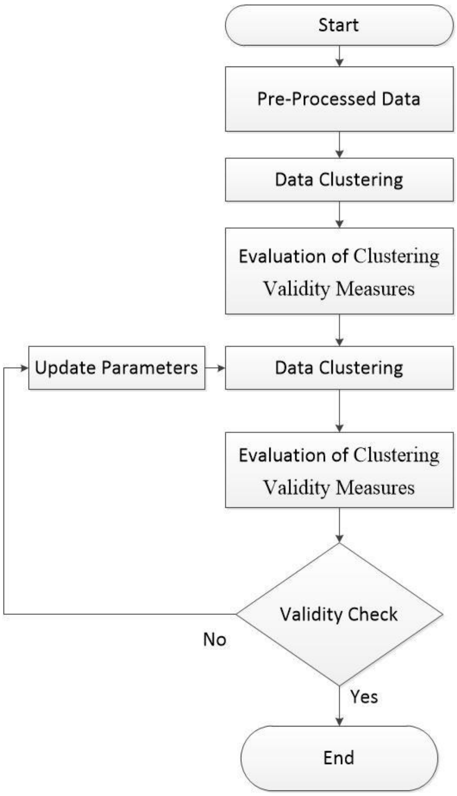

Figure 4. Widely adopted clustering mechanism.

However, there are new ways of optimally and automatically selecting the number of clusters based on the load profile data itself. In a recent work by [109], the authors noted that the traditional K-Means method does not capture peaks and truffles of the original load profile while clustering. Hence, a new technique called Slope-based shape clustering was formulated, resulting in better clustering with improved computation efficiency. Another major setback is the rigidity of the process; the number of clusters remains the same throughout the iteration. In contrast, it may turn out that more or fewer clusters would fit the data better. This led to the use of novel clustering algorithms such as ISODATA and affinity propagation akin to K-Means. These clustering techniques automatically adjust the number of clusters during iteration.

The K-Means algorithm was also identified as impracticable when the number of customers or dimensions is high [23][110]. Noting that the algorithm's accuracy fades away with increment in the number of customers, as it is susceptible to noise and outliers [108]. This assertion was proved in [22], that K-means, K-medoids and SOM were used to cluster a relatively large sample size. From the results obtained, SOM outshone K-means by giving a consistently lower DB index overall across varying numbers of clusters. [19] therefore suggested the use of K-Means when the number of entities concerned is minimal. Fuzzy c-means clustering is another similar technique to K-Means clustering. The significant difference with this algorithm is: all data points belong to all clusters to some degree, such that the degree of membership of each data point from all clusters is equal to one [103]. This technique also has the challenge of initial selection of the number of clusters, K. However, when compared with K-Means clustering, fuzzy c-means has a long process because both the centroids and the degree of membership of each data point are updated at every iteration. Hence, it is hardly employed in works of literature. Also, its inferiority to K-Means has been proved in several works [93][102].

b) Hierarchical clustering

Hierarchical Clustering is a more deterministic and flexible approach to clustering data objects as it does not mandate the predefinition of the number of clusters [73]. It is considered the second most crucial clustering technique after partitional clustering, and it is subdivided into two (2) major types, named Agglomerative clustering and Divisive clustering methods [14]. Agglomerative Clustering starts by taking singleton clusters, that is, making every data point a cluster point. The singleton clusters are gradually merged to form a build-up hierarchy of clusters using a dissimilarity matrix until the final maximal cluster is obtained in a tree-like structure. The optimal number of clusters is selected from the dendrogram drawn by obtaining the most prominent vertical distance that does not intersect any clusters. A horizontal line is drawn at the extreme ends of the point; the number of vertical lines passing through the horizontal line gives the optimal number of clusters [111]. The Divisive Clustering method uses the top-down approach in which a hierarchy of clusters is generated by continuously splitting a huge macro-cluster into 2 two (2) groups [73]. The tree-like structure formed after clustering using any of the approaches is called a dendrogram. [71] noted that the Divisive Hierarchical clustering is not always used in clustering consumption data in a power system due to the algorithm's complexity and high computational time. Hence, the agglomerative clustering methods were employed in works of literature that employed hierarchical clustering. There are different kinds of agglomerative clustering methods. These methods differ majorly in the similarity measures they employ, and they include the single link (nearest neighbour), complete link (diameter), group average (average link), centroid similarity, and Ward's criterion (minimum variance) [73]. One major shortcoming found with the Hierarchical clustering technique is that it is not flexible. Once a dendrogram is merged or split, it cannot be undone. Also, [37] noted that Hierarchical clustering is not sturdy towards outliers as they may emerge as additional clusters. Hence, they produce less quality clustering when compared to partitional clustering algorithms such as K-means.

c) Self-Organising Map (SOM)

Self-Organising Map (SOM) is an unsupervised clustering algorithm that uses artificial neural networks and adopts a two (2) dimensional grid [77]. Each neuron is connected to the neighbouring neurons, hence, the Euclidean distance between the input vector and the neuron is updated whenever the winner neuron is computed. This process is repeated for all input data, and at the end, the most similar data are allocated together or closely [14]. However, some initial parameters such as learning rate, number of neurons and number of epochs need to be initialised before clustering. The SOM algorithm can map complex and non-linear relationships into lower-dimensional data; hence, it is often used to find patterns between electricity consumption and demographics. It is well known for its outstanding resilience to outliers and missing values, which helps in reducing noise in data. Hence, in [77] and [22], where the sample size population is exceptionally large, SOM outshone K-Means. SOM algorithm is also good at creating data abstraction rather than removing missing values from consumption data [34]. Due to this uniqueness, SOM is primarily used in indirect clustering much more than direct clustering, [89] noted that the algorithm is much more effective in second-order clustering works.

Table 4. Clustering algorithms adopted by literature.

SN	Clustering Algorithm	References	SN	Clustering Algorithm	References
1	Iterative Self-Organizing Data-Analysis Technique (ISODATA)	[44]	**14**	Adaptive K-Means	[20]
2	Hierarchical, K-Means, fuzzy c-means, and two-stage clustering	[103]	**15**	SOM and K-Means	[77][78] [29] [89]
3	Hierarchical and Fuzzy C Means	[66]	**16**	K-means and Hierarchical Clustering	[50]
4	Fuzzy C-Means Clustering, Markov Model	[111]	**17**	The Density-based Spatial Clustering of Application with Noise (DBSCAN)	[68]
5	K-Means	[9] [13] [14][15][30][36] [37] [40] [48] [51] [53] [61] [65] [80][82][107][112][113][114] [115][116]	**18**	K-means, Complete-link (CL), Average-link (AL), Ward ́s-link (WL), Normalized Cut algorithm (NC).	[45]
6	K-means, normalised N-Cut, Pairwise Constrained (PC k-means) and Metric Pairwise Constrained (MPC k-means).	[41]	**19**	K-Means, Fuzzy C-Means, Hierarchical Complete linkage, And Hierarchical Ward’s Methods	[93]
7	SOM + Hierarchical	[78][89]	**20**	K-means and Fuzzy Logic Clustering	[106]
8	SOM +K-Means++	[94]	**21**	Hierarchical Clustering	[38] [42] [81] [117]
9	K-Means, Hierarchical and Dirichlet Process Mixture Model (DPMM)	[19]	**22**	Slope-based Shape Cluster Method	[109]
10	DBSCAN and K-Means	[23][105]	**23**	Affinity Propagation	[110]
11	Fast Search and Find of Density Peaks (CFSFDP)	[47]	**24**	Follow up Leader	[64]
12	K-means and K-medoids	[10]	**25**	Gaussian Mixture Model (GMM) and Hierarchical Clustering	[69]
13	Non-negative K-SVD algorithm	[86]	**26**	CVVM	[79]

4.3.2. Clustering parameters

Aside from the need to careful select the clustering algorithms for load profiling, some salient parameters are also pivotal. Among them are the distance metrics and clustering validation measures. Distance metrics is used for measuring the distance between a pair sequence when clustering [73]. The distance metric is preselected before clustering, while the validation measure is used as a similarity validation index for every formed cluster.

Distance Metrics

A distance metric takes a pair of sequences and returns a real number which denotes the distance between the given sequences [73]; this is very significant in clustering algorithms. For instance, as illustrated in the K-Means algorithm, the closest centroid is computed iteratively based on the data points' proximities (distance calculations) [73]. Different kinds of proximity measures can be adopted: Manhattan distance, Euclidean distance, Hamming distance, Cosine Similarity and many more [103]. The choice of the proximity measure adopted during clustering significantly impacts quality of clustering [61]. Thus, in a comparative analysis carried out by [61] on the use of Euclidean distance, Manhattan distance and Dynamic Time Warping when using the K-Means algorithm on a daily load profile. The clustering result inferred that Euclidean distance produces the most consistent results among others. Among all the proximity measures in literature, Euclidean distance is the most used [4][44][73]. However, a recent similarity measure was adopted by [109], where the curve slope method was used. The result obtained inferred the new method to be more efficient. Also, in a recent comparison by [67] on the effectiveness of the Pearson Correlation Coefficient and Euclidean distance in the clustering daily load profiles using K-Mean clustering, the former outperformed the latter.

Clustering Validation Measures

Clustering validation measures are ways of evaluating the results of a clustering algorithm [14], a good cluster is known to present records with high similarity among them [41]. According to [73], clustering validation measures can either be external or internal. External validation measures use external information not present in the data to evaluate the extent of a cluster. However, external clustering validation was rarely used in previous works of clustering electricity data [19].

The internal validation evaluates the goodness of the clustering structure without respect to external information. Internal clustering validation can be used to choose the number of clusters and the best clustering algorithm without external information. Hence, many internal clustering validation criteria have been developed and employed by works of literature to validate the effectiveness of the number of clusters preselected. Also considered is the compactness of each cluster and how separated the clusters are from one another after clustering. Among them are Calinski–Harabasz Index (CHI) [29], Silhouette Coefficient (S score) [13][103], the Scatter Index (SI) [19], Clustering Dispersion Indicator (CDI)[19], Davies-Bouldin index (DBI) [20] etc. Furthermore, as inferred from [81], cluster validity indexes can also be used to validate the best alphabets numbers as in the case of dimension reduction using SAX. Discussed below are the five (5) most used clustering validation indexes for load profiling.

(i) The Davies–Bouldin index (DBI): To evaluate the DBI of a clustering output, for every cluster C, the similarities between C and other clusters are calculated. The highest value in terms of proximity is apportioned to C as its cluster similarity. The DBI is obtained by finding the average of all the cluster similarities. A smaller DBI value indicates that the clusters are more distinct, hence better clustering [73].

(ii) The Silhouette Index (SI): The Silhouette calculation is done via the equation in Table 5. to validate the clustering performance based on the pairwise difference between and within-cluster distances [73]. Where a(i) is the average distance to other objects in the same cluster and b(i) is the average distance to objects of the nearest cluster. In this case, the optimal cluster number is determined by maximizing the value of this index. However, the index value ranges between -1 and 1. Hence, to evaluate the quality of a clustering technique, one can compute the average silhouette coefficient of all points in the dataset.

(iii) The Dunn Index (DI): The Dunn Index is the quotient of the d_{min} and $d_{\max}$ d_{min} refers to the minimum distance between points of different clusters, while $d_{\max}$is the largest distance between two (2) distinct points within a cluster, the maximum across the clusters is selected as $d_{\max}$ The number of clusters that give the highest DI infers a better clustering[118].

(iv) Mean Index Adequacy (MIA): This refers to the average of all the distances between the objects in the clusters and corresponding centroid. The minimum value from a varying number of clusters infers the best clustering result.

(v) Clustering Dispersion Indicator (CDI): This refers to the ratio of the mean intra-set distance between the patterns in the same cluster and the inter-set distance between the cluster's centroids [103]. The maximum value obtained is used to infer a better clustering result.

Illustrated in Table 5 are the various clustering validity indices employed by works of literature, coupled with their respective mathematical definitions and rules. Many authors used more than a clustering algorithm and subjected them to as many validity indices as possible. In the bid to obtain the clustering algorithm with the best performance among all [15][19][41][45]. Some works only made use of the validity indices as guidance in the preselection of clustering number or reduction technique [9][13][20][29][40][65][66][89][105]. Among the clustering indices used are some novel and unpopular clustering indices, such as the Xien-Ben cluster validity index (XB) and Point Symmetry index (PSI).

Table 5. Clustering validity measures employed by literatures.

Clustering Validity Indices (CVI)	Literature	Definition	Rule
Clustering Dispersion Indicator (CDI)	[9][15] [19] [40] [79] [110]	$CDI = \frac{D_{max}}{D_{max}} \sum_{i=1}^{k} \left(\sum_{j=1}^{k} d(r^{(i)}, r^{(j)}) \right)^{-1}$	Maximum
Davies-Bouldin index (DBI)	[9] [19] [20][22] [36] [40] [45] [79] [81][89] [93] [110]	$DBI = \frac{1}{K} \sum_{x=1}^{k} \left(\frac{d'(L^{(i)} + d'(L^{(i)}))}{d(r^{(i)}, r^{(i)})} \right)$	Minimum
Dunn index (DI)	[19] [23] [36] [41] [45] [67] [69] [103]	$DI = \frac{(d_{min})}{(d_{max})}$	Maximum
Silhouette Index (SI)	[13][29] [36] [40] [45] [53] [65] [103] [110] [103]	$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$	Maximum
Mean Index Adequacy (MIA)	[9][19] [79] [81]	$MIA = \sqrt{\frac{\sum_{i=1}^{K} d(r^{(i)}, L^{(i)})}{K}}$	Minimum

K = Total number of Clusters, $L^{(i)}$ = Set of Objects in cluster I, $r^{(i)}$ = centre of cluster i, d = Sum of the distance between objects in the cluster and the cluster centre, $d'(L^{(i)})$ = Geometric mean of the inter- distance between objects in $L^{(i)}$, $d(r^{(i)}, r^{(j)})$ = Distance between centres of cluster i and j, D_{max} = Maximum distances between the cluster centres, D_{min} = Minimum distances between the cluster centres, d_{min} = minimal distance between points of different clusters, $d_{\max}$ = the largest within-cluster distance.

Computational Cost of Clustering

Big data analysis comes with high computational costs [11]. However, it has been observed that computational cost during clustering is dependent on the algorithm adopted. Thus, to minimise this cost, works of literature have been careful in choosing a clustering algorithm for load profiling. During an analysis carried out by [75] on the generation of load profiles from customers with automatic meter readings. He observed a trade-off between clustering efficiency and computational cost. On subjecting energy data to the three (3) most used clustering algorithms: K-Means, Hierarchical and Fuzzy-c clustering. The results obtained indicate that hierarchical clustering takes less and consistent processing time over others, regardless of the number of clusters, but it was inefficient in reducing the Mean Absolute Error. Whereas K-Means clustering was efficient in reducing the Mean Absolute Error but the processing time increases as the number of clusters increases. Fuzzy-c clustering, however, took twice the time required by K-Mean clustering, but the result is less efficient to that of K-Means. Hence, [73] noted that K-means is only more computationally efficient over other techniques when the number of clusters is minimal.

4.3.3. Post-Clustering

After clustering consumption data, the classifications need to be properly studied for efficient interpretation of the classes obtained. A class among the clusters may represent outliers. Hence, the user must technically interpret the result. However, the classes obtained can be interpreted in various ways; each class may represent specific seasonal, economic, technical attributes etc. [72]. A Representative Load Profile (RLP) is assigned to each class based on any of these attributes. Each RLP may be formulated by evaluating the median of each profile in a class or the extraction of the centroids [30]. However, there are bound to be many fluctuations in the averaged load profiles, making it difficult to extract patterns [71]. These noises can be minimised by introducing a curve smoothening technique to smoothen the data. Polynomial curve fitting and moving average smoothing techniques are outstanding techniques often employed. The polynomial curve fitting technique requires the use of higher degree polynomials to handle higher variations. However, according to Runge’s phenomenon, high degree polynomials are generally unsuitable for interpolation with equidistant nodes. Hence the average smoothening technique has been widely adopted in current scenarios,due to its simplicity and accuracy [26][71]. However, results from [71] show that smoothing before profiling tends to remove peak energy demands that can be found in the raw data. However, he noted that the effect would be minimal for applications such as Demand Side Management (DSM) programs, as peaks at this level are for a very short duration.

4.3.4. Impact of data resolution on clustering

The smart-meter determines the resolution of the acquired dataset. Research has shown that the data resolution of a smart meter data affects the clustering quality [31][61]. In a research work [19], the impact of data resolution on clustering results was examined on K-Means, Hierarchical and Dirichlet Process Mixture Model (DPMM) algorithms by varying the raw data resolution and subjecting the results obtained to several validity measures. The research result affirmed that the consumption data’s resolution substantially affect the cluster quality and cluster membership consistency. The authors then advised that to obtain accurate and sufficient differences between consumers, the frequency of the data should be at least 30 minutes.

However, [51] established a different fact; they noted that the number of clusters selected has a more significant effect on the clustering result than the resolution of the data. Their work aggregated

1-minute consumption data into Typical Weekly Profiles (TWPs) and Typical Daily Profiles (TDPs). Using DBI and the DI as indicators, no improvement was seen in the clustering result when the time resolution was increased. The authors thus opined that the choice of data resolution is application dependent; in their works, they cited TWP to be more efficient for revealing consumers that have equipment turned on than TDP, a knowledge that is very instrumental to applications such as the implementation Time-of-Use tariffs.

5. Discussion

Our findings show a gradual replacement of statistical methods by intelligent techniques during the pre-processing and profiling of smart meter data across works of literature. The often-overlooked non-linearity of data feature bolstered the need for replacements, methodised in statistical learning methods. This nuance was unmasked at every stage of pre-processing smart meter data, i.e., outlier detection, missing data replacement and data reduction, discussed in this article. Data imputation, conventionally carried out using interpolation and regression techniques, is now optimized using non-linear algorithms. Also, the compulsory data reduction often done for smart meter data analytics are now implemented using deep learning methods. Advancement was also found in the detection of outliers; among the various mining techniques.

Clustering technique as an unsupervised learning technique has been found resourceful; the use of DBCAN is more prominent in recent detecting outliers, considering the arduity and inaccuracy that accompany other non-intelligent techniques. These advances not only improve the quality of the results obtained but also reduce the computational cost.

To identify hidden patterns, using aggregated data to profile will lead to the loss of some information. Hence, the use of clustering, among various clustering techniques, K-Means, Hierarchical, fuzzy c-means, Follow-up ladder, and Self-Organising Map (SOM) were considered in the review due to the simplicity and accuracy they command [15][31][53][73][104]. However, as shown in Table 6, from works of literature, K-Means has been proven to be the best clustering algorithm using various clustering validity indices [120]. The technique maintains superiority over other techniques. This assertion was validated across the considered works of literature using synthetic and actual smart meter data.

6. Conclusion

Data mining techniques were found to be very instrumental in the pre-processing (detecting outliers and the treatment of missing data) of consumption data and its profiling. However, the nature of the consumption data plays a significant role in the choice of the DM technique adopted at every phase of load analysis. The review shows that results can be more profound and accurate when the raw consumption data is carefully subjected to novel reduction techniques such as feature extraction, depending on the intrinsic nature of the dataset. From the results obtained, the comparison was done on standard existing clustering algorithms employed in several works towards profiling streams of consumption data. It is concluded that the K-means algorithm, when guided by appropriate clustering validity indexes, presents better results compared with other algorithms in most cases. Albeit, the technique’s efficiency is hampered by setbacks such as the need to predefine the number of clusters. This has engendered a myriad of novel techniques that are yet to be widely adopted but has been validated by a few works of literature. Therefore, further validation and modification is imperative for a robust K-Means, for better smart meter data profiling.

Table 6. Clustering techniques comparison in works of literature using CVI.

SN	Clustering Validity Indices (CVI)	Literature	Clustering Algorithms or Reduction Techniques Compared	Best Performed Clustering Algorithm
1	DI, DBI, Root-mean-square standard error (RMSSTD), R-squared index (RSI), the SD validity index, SI, Xien-Ben cluster validity index (XB), Squared Error index (SQE), and the Point Symmetry index (PSI)	[45]	K-means, Complete-link (CL), Average-link (AL), Ward ́s-link (WL), Normalized Cut algorithm (NC).	K-Means
2	CDI, DBI, DI MDI, MIA, SI, and Variance Ratio Criterion (VRC).	[19]	K-Means, Hierarchical and Dirichlet Process Mixture Model (DPMM)	K-Means
3	SD validity index, PSI, DBI, XBI, DI, the Normalized Hubert Statistic (NH)	[41]	K-means, normalised N-Cut, Pairwise Constrained (PC k-means) and Metric Pairwise Constrained (MPC k-means)	K-Means
4	SI	[102]	Hierarchical, K-Means, fuzzy c-means, and two-stage clustering	K-Means,
		[83]	Hierarchical and K-Means	K-Means
5	CDI, MDI, DBI and MIA	[79]	CVMM, Hierarchical (Average), Hierarchical (Weighted), Hierarchical (Complete), K-means, Hierarchical (Ward), Gaussian mixture model (GMM)	CVMM
6	DBI	[22]	K-Means, k-medoid and SOM	SOM
		[93]	K-Means, Fuzzy C-Means, Hierarchical Complete linkage, And Hierarchical Ward’s Methods	Hierarchical Complete linkage
7	DBI, CHI and SI	[65]	PCA + K-Means, Locally Linear Embedding (LLE) + K-Means, PAA + SAX + K-Means	PAA + SAX + K-Means

8	MIA, CDI, DBI	[9]	Variance score, Laplacian Score, Sparse K-Means score (SK-Means), Proposed method	Proposed method + K-Means
9	Adjusted Rand Index (ARI)	[77]	K-Means and SOM	SOM
10	SI, CHI, DBI, DI, Ratio of the within-cluster sum of squares to between-cluster variation (WCBCR) and CDI	[92]	Affinity Propagation, k-mean, k-medoids and spectral clustering	Affinity Propagation
11	Elbow method, SI, DI	[120]	k-means, fuzzy k-means, agglomerative hierarchical	K-Means
12	DI	[23] [67]	DBSCAN + K-Means, Gaussian mixture model (GMM) clustering algorithm, K-Means.	DBSCAN + K-Means.
		[69]	Partitioning Around Medoids (PAM) clustering, Hierarchical clustering.	Hierarchical clustering.

[45] established the superiority of K-Means in a classification work carried out on 208 medium voltage (MV) electricity consumers in a smart grid environment. The result obtained from K-Means were compared with four (4) other different clustering algorithms (Complete-link (CL); Average-link (AL); Ward ́s-link (WL); Normalized Cut (NC)), using twelve (12) clustering validity indices. The 12 indices all asserted K-Means as the best as it gave the best partition. K-Means also showed the best clustering in [41] when compared with the other three (3) clustering algorithms (normalised N-Cut, Pairwise Constrained (PC K-means), and Metric Pairwise Constrained (MPC K-means)) using eight (8) validity indices. Furthermore, during the generation of a Virtual Load Profile (VLP) for consumers with Non-Automated Meter Reading (non-AMR) by [103] from Typical Load Profile (TLP) data obtained from clustered AMR consumption data, clustering performance from Hierarchical, K-Means, fuzzy c-means, and two-stage clustering were compared. K-Means clustering also showed a better Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Silhouette value in each cluster set.

However, K-means clustering is encumbered with some significant setbacks that are being mitigated by works of literature to sustain the quality result the method commands. One of the biggest shortcomings with traditional K-Means algorithms is that the number of clusters that needs to be pre-defined before the clustering processes, a decision that directly influences the identification of the typical load profiles [20][67]. This lag engendered some novel methods and packages such as the replication analysis, the lower bound technique, NbClust and many more [82][121]. These updates led to the modification of the K-Means algorithm, as in K-Means++, adaptive K-Means, ISODATA [94]. However, the efficiencies of these novel algorithms cannot be affirmed generally to be more than the traditional K-Means because they have only been employed in a few works of literature [44][110].

Acknowledgements

This research was funded by TETFund Research Fund Grant 2019 (TETFund/DR&D-CE/NRF/2019).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Kabalci Y (2016) A survey on smart metering and smart grid communication. *Renew Sustain Energy Rev* 57: 302–318. doi: 10.1016/j.rser.2015.12.114.
2. Dileep G (2020) A survey on smart grid technologies and applications. *Renew Energy* 146: 2589–2625. doi: 10.1016/j.renene.2019.08.092.
3. Grolinger G, Capretz MAM, Seewald L (2016) Energy consumption prediction with big data: Balancing prediction accuracy and computational resources. *Proc. - 2016 IEEE Int Congr. Big Data, BigData Congr* 90: 157–164. doi: 10.1109/BigDataCongress.2016.27.
4. Yang T, Ren M, Zhou K (2017) Identifying household electricity consumption patterns: A case study of Kunshan, China. *Renew Sustain Energy Rev* 91: 861–868. doi: 10.1016/j.rser.2018.04.037.
5. Garcia FD, Marafao FP, De Souza WA, Da Silva LCP (2017) Power Metering: History and Future Trends. *IEEE Green Technol Conf*, 26–33. doi: 10.1109/GreenTech.2017.10.
6. Guerrero-Prado JS, Alfonso-Morales W, Caicedo-Bravo E, et al. (2020) The power of big data and data analytics for AMI data: A case study. *Sensors (Switzerland)* 20: 1–27. doi: 10.3390/s20113289.
7. Wang Y, Qiu H, Tu Y, et al. (2018) A Review of Smart Metering for Future Chinese Grids. *Energy Procedia* 152: 1194–1199. doi: 10.1016/j.egypro.2018.09.158.
8. Quilumba FL, Lee WJ, Huang H, et al. (2014) An overview of AMI data preprocessing to enhance the performance of load forecasting. *2014 IEEE Ind Appl Soc Annu Meet IAS 2014*, 1–7. doi: 10.1109/IAS.2014.6978369.
9. Khan I, Huang JZ, Masud A, et al. (2016) Segmentation of Factories on Electricity Consumption Behaviors Using Load Profile Data. *IEEE Access* 4: 8394–8406. doi: 10.1109/ACCESS.2016.2619898.
10. Ali U, Buccella C, Cecati C (2016) Households electricity consumption analysis with data mining techniques. *IECON Proc Industrial Electron Conf*, 3966–3971. doi: 10.1109/IECON.2016.7793118.
11. Wang Y, Chen Q, Hong T, et al. (2019) Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Trans Smart Grid* 10: 3125–3148. doi: 10.1109/TSG.2018.2818167.
12. Tshikomba SC, Estrice M, Ojo E, et al. (2020) Curbing Electricity Theft Using Wireless Technique with Communication Constraints. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1–6. doi: 10.1109/icABCD49160.2020.9183812.

13. Zala HN, Abhyankar AR (2014) A Novel Approach to Design Time Of Use Tariff using Load Profiling and Decomposition. *2014 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, 1–6.
14. Cembranel SS, Lezama F, Soares J, et al. (2019) A Short Review on Data Mining Techniques for Electricity Customers Characterization. *2019 IEEE Power and Energy Society*, 194–199.
15. Colley D, Mahmoudi N, Eghbal D, et al. (2014) Queensland Load Profiling by Using Clustering Techniques. *2014 Australasian Universities Power Engineering Conference (AUPEC)*, 1–6.
16. Martinez-Pabon M, Eveleigh T, Tanju B (2017) Smart Meter Data Analytics for Optimal Customer Selection in Demand Response Programs. *Energy Procedia* 107: 49–59. doi: 10.1016/j.egypro.2016.12.128.
17. Liao SH, Chu PH, Hsiao PY (2012) Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Syst Appl* 39: 11303–11311. doi: 10.1016/j.eswa.2012.02.063.
18. Khan ZA, Jayaweera D (2017) Approach for smart meter load profiling in Monte Carlo simulation applications. *IET Gener Transm Dis* 11: 1856–1864. doi: 10.1049/iet-gtd.2016.2084.
19. Granell R, Axon CJ, Wallom DCH (2015) Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles. *IEEE Trans POWER Syst* 30: 3217–3224. doi: 10.1109/TPWRS.2014.2377213.
20. Lu S, Lin G, Liu H, et al. (2019) A Weekly Load Data Mining Approach Based on Hidden Markov Model. *IEEE Access* 7: 34609–34619. doi: 10.1109/ACCESS.2019.2901197.
21. Gautam A (2015) Load Profile Determination of Consumer in a Smart Grid by Using Matrix Based Approach. *2015 Annu IEEE India Conf*, 1–6.
22. Mcloughlin F, Duffy A, Conlon M (2015) A clustering approach to domestic electricity load profile characterisation using smart metering data. *Appl Energy* 141: 190–199. doi: 10.1016/j.apenergy.2014.12.039.
23. Liu X, Ding Y, Tang H, et al. (2021) A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy Build* 231: 110601. doi: 10.1016/j.enbuild.2020.110601.
24. Kelly J, Knottenbelt W (2015) The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Nature* 2: 1–14. doi: 10.1038/sdata.2015.7.
25. Manembu P (2015) Missing Data Solution of Electricity Consumption based on Lagrange Interpolation Case Study : IntelligEnSia data monitoring. *The 5th International Conference on Electrical Engineering and Informatics 2015*, 511–516.
26. Khan ZA, Jayaweera D (2018) Approach for Forecasting Smart Customer Demand With Significant Energy Demand Variability. *2018 1st International Conference on Power, Energy and Smart Grid (ICPESG)*, 1–5. doi: 10.1109/ICPESG.2018.8384528.
27. Zhao Y, Zhang C, Zhang Y, et al. (2020) A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy Built Environ* 1: 149–164. doi: 10.1016/j.enbenv.2019.11.003.
28. Fan C, Xiao F, Li Z, et al. (2018) Unsupervised data analytics in mining big building operational data for energy efficiency enhancement : A review. *Energy Build* 159: 296–308. doi: 10.1016/j.enbuild.2017.11.008.
29. Choksi KA, Jain S, Pindoriya NM (2020) Feature based clustering technique for investigation of domestic load profiles and probabilistic variation assessment: Smart meter dataset. *Sustain Energy Grids Networks* 22: 1–11. doi: 10.1016/j.segan.2020.100346.

30. Zhan S, Liu Z, Chong A, et al. (2020) Building categorization revisited : A clustering-based approach to using smart meter data for building energy benchmarking. *Appl Energy* 269. doi: 10.1016/j.apenergy.2020.114920.
31. Wang Y, Chen Q, Kang C, et al. (2015) Load Profiling and Its Application to Demand Response : A Review. *Tsinghua Sci Technol* 20: 117–129. doi: 10.1109/TST.2015.7085625.
32. Funde NA, Dhabu MM, Paramasivam A, et al. (2019) Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data. *Sustain Cities Soc* 46. doi: 10.1016/j.scs.2018.12.043.
33. Liberati A, Altman DG, Tetzlaff J, et al. (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol* 62: e1–e34.
34. Kolehmainen M, Mononen M, Niska H (2015) A Data Mining Approach for Producing Small Area Statistics-based Load Profiles for Distribution Network Planning. *2015 IEEE International Conference on Industrial Technology (ICIT)*, 1236–1240. doi: 10.1109/ICIT.2015.7125266.
35. Manembu P, Engineering I, Nielsen PS, et al. (2018) Multi-grained Household Load Profile Analysis using Smart Meter Data : The Case of Indonesia. *2018 2nd Borneo International Conference on Applied Mathematics and Engineering (BICAME)*, 213–217. doi: 10.1109/BICAME45512.2018.1570503357.
36. Viegas JL, Vieira SM, Mel ćio R, et al. (2016) Classification of new electricity customers based on surveys and smart metering data. *Energy* 107: 804–817. doi: 10.1016/j.energy.2016.04.065.
37. Azad SA, Ali ABMS, Wolfs P (2014) Identification of Typical Load Profiles using K -Means Clustering Algorithm. *Asia-Pacific World Congress on Computer Science and Engineering*, 1–6. doi: 10.1109/APWCCSE.2014.7053855.
38. Gunsay M, Bilir C, Poyrazoglu G (2020) Load Profile Segmentation for Electricity Market Settlement. *2020 17th International Conference on the European Energy Market (EEM)*, 1–5. doi: 10.1109/EEM49802.2020.9221889.
39. Wang Y, Chen Q, Kang C, et al. (2016) Residential smart meter data compression and pattern extraction via non-negative K-SVD. *IEEE Power Energy Soc Gen Meet*, 1–5. doi: 10.1109/PESGM.2016.7741464.
40. Tureczek A, Nielsen PS, Madsen H (2018) Electricity consumption clustering using smart meter data. *Energies* 11: 1–18. doi: 10.3390/en11040859.
41. Ramos S, Duarte JM, Duarte FJ, et al. (2013) A data mining framework for electric load profiling. *2013 IEEE PES Conf Innov Smart Grid Technol ISGT LA*, 1–6. doi: 10.1109/ISGT-LA.2013.6554489.
42. Gouveia JP, Seixas J, Mestre A (2017) Daily electricity consumption profiles from smart meters - Proxies of behavior for space heating and cooling. *Energy* 141: 108–122. doi: 10.1016/j.energy.2017.09.049.
43. Su T, Shi Y, Yu J, et al. (2021) Nonlinear compensation algorithm for multidimensional temporal data: A missing value imputation for the power grid applications. *Knowledge-Based Syst* 215: 106743. doi: 10.1016/j.knosys.2021.106743.
44. Mutanen A, Ruska M, Repo S, et al. (2011) Customer Classification and Load Profiling Method for Distribution Systems. *IEEE Trans Power Deliv* 26: 1755–1763.
45. Ramos S, Duarte JMM, Soares J, et al. (2012) Typical load profiles in the smart grid context a clustering methods comparison. *IEEE Power and Energy Society General Meeting*, 1–8. doi: 10.1109/PESGM.2012.6345565.

46. Mononen M, Saarenp ää J, Johansson M, et al. (2014) Data-driven method for providing feedback to households on electricity consumption. *2014 IEEE 9th International Conference on Intelligent Sensors, Sensor Networks and Information Processing,* 1–6. doi: 10.1109/ISSNIP.2014.6827661.
47. Wang Y, Member S, Chen Q, et al. (2016) Clustering of Electricity Consumption Behavior Dynamics toward Big Data Applications. *IEEE Trans Smart Grid* 3053: 1–11. doi: 10.1109/TSG.2016.2548565.
48. Trotta G (2020) An empirical analysis of domestic electricity load profiles: Who consumes how much and when? *Appl Energy* 275: 115399. doi: 10.1016/j.apenergy.2020.115399.
49. Kim N, Kim M, Choi JK (2018) LSTM Based Short-term Electricity Consumption Forecast with Daily Load Profile Sequences. *2018 IEEE 7th Global Conference on Consumer Electronics, GCCE*, 834–835. doi: 10.1109/GCCE.2018.8574484.
50. Jain S, Choksi KA, Pindoriya NM (2019) Rule-based classification of energy theft and anomalies in consumers load demand profile. *IET Smart Grid* 2: 612–624. doi: 10.1049/iet-stg.2019.0081.
51. Frost AE, Azaza M, Li H, et al. (2017) Patterns and temporal resolution in commercial and industrial typical load profiles. *8th International Conference on Applied Energy – ICAE2016*, 105: 2684–2689. doi: 10.1016/j.egypro.2017.03.775.
52. Ndiaye D, Gabriel K (2011) Principal component analysis of the electricity consumption in residential dwellings. *Energy Build* 43: 446–453. doi: 10.1016/j.enbuild.2010.10.008.
53. Yilmaz S, Chambers J, Patel MK (2019) Comparison of clustering approaches for domestic electricity load pro fi le characterisation - Implications for demand side management. *Energy* 180: 665–677. doi: 10.1016/j.energy.2019.05.124.
54. Xu C, Chen H (2020) A hybrid data mining approach for anomaly detection and evaluation in residential buildings energy data. *Energy Build* 215: 109864. doi: 10.1016/j.enbuild.2020.109864.
55. Wang MC, Tsai CF, Lin WC (2021) Towards missing electric power data imputation for energy management systems. *Expert Syst Appl* 174: 114743. doi: 10.1016/j.eswa.2021.114743.
56. Mateos G, Giannakis GB (2012) Spatiotemporal Load Curve Data Cleansing and Imputation via Sparsity and Low Rank. *IEEE Smard Grid Communication,* 653–658.
57. Peppanen J, Zhang X, Grijalva S, et al. (2016) Handling Bad or Missing Smart Meter Data through Advanced Data Imputation. *2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT),* 1–5. doi: 10.1109/ISGT.2016.7781213.
58. Tkachenko R, Mishchuk O, Izonin I, et al. (2019) A non-iterative neural-like framework for missing data imputation. *Procedia Comput Sci* 155: 319–326. doi: 10.1016/j.procs.2019.08.046.
59. Che Z, Purushotham S, Cho K, et al. (2018) Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci Rep* 8: 1–12. doi: 10.1038/s41598-018-24271-9.
60. Ma J, Cheng JCP, Jiang F, et al. (2016) A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy Build* 216: 109941. doi: 10.1016/j.enbuild.2020.109941.
61. Bourdeau M, Basset P, Beauchene S, et al. (2021) Classification of daily electric load profiles of non-residential buildings. *Energy Build* 233: 110670. doi: 10.1016/j.enbuild.2020.110670.
62. Jurj DI, Micu DD, Czumbil L, et al. (2020) Analysis of Data Cleaning Techniques for Electrical Energy Consumption of a Public Building. *UPEC 2020 - 2020 55th International Universities Power Engineering Conference, Proceedings,* 1–6. doi: 10.1109/UPEC49904.2020.9209781.

63. Zheng K, Chen Q, Wang Y, et al. (2019) A Novel Combined Data-Driven Approach for Electricity Theft Detection. *IEEE Trans Ind Informatics* 15: 1809–1819. doi: 10.1109/TII.2018.2873814.
64. Piscitelli MS, Brandi S, Capozzoli A (2019) Recognition and classification of typical load profiles in buildings with non- intrusive learning approach. *Appl Energy* 255: 113727. doi: 10.1016/j.apenergy.2019.113727.
65. Shi Y, Yu TAO, Liu Q, et al. (2020) An approach of electrical load profile analysis based on time series data mining. *IEEE Access* 8: 209915–209925. doi: 10.1109/ACCESS.2020.3019698.
66. Lin X, Wu W, Zeng B, et al. (2018) Analysis of large-scale electricity load profile using clustering method. *ICNSC 2018 - 15th IEEE Int. Conf. Networking, Sens. Control*, 1–5. doi: 10.1109/ICNSC.2018.8361335.
67. Ma Z, Yan R, Nord N (2017) A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. *Energy* 134: 90–102. doi: 10.1016/j.energy.2017.05.191.
68. Yang J, Member S, Zhao J, et al. (2018) A Model of Customizing Electricity Retail Prices Based on Load Profile Clustering Analysis. *IEEE Trans Smart Grid* 10: 3374–3386. doi: 10.1109/TSG.2018.2825335.
69. Li K, Ma Z, Robinson D, et al. (2018) Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering. *Appl Energy* 231: 331–342. doi: 10.1016/j.apenergy.2018.09.050.
70. Celis S, Giraldo LF, De Oliveira-De Jesus P, et al. (2018) A Clustering Approach for Domestic Smart Metering Data Preprocessing. *2018 IEEE ANDESCON, ANDESCON 2018 - Conf. Proc.*, 1–3. doi: 10.1109/ANDESCON.2018.8564597.
71. Khan ZA, Jayaweera D, Alvarez-alvarado MS (2018) A novel approach for load profiling in smart power grids using smart meter data. *Electr Power Syst Res* 165: 191–198. doi: 10.1016/j.epsr.2018.09.013.
72. Chicco G (2012) Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy* 42: 68–80. doi: 10.1016/j.energy.2011.12.031.
73. Aggarwal CC, Reddy CK (2014) *DATA Algorithms and Applications*. Minnesota, USA.
74. Zucker G, Habib U, Blöchle M, et al. (2015) Sanitation and analysis of operation data in energy systems. *Energies* 8: 12776–12794. doi: 10.3390/en81112337.
75. Kim YI, Ko JM, Choi SH (2011) Methods for generating TLPs (Typical Load Profiles) for smart grid-based energy programs. *IEEE SSCI 2011 - Symp. Ser. Comput. Intell. - CIASG 2011 2011 IEEE Symp Comput Intell Appl Smart Grid*, 49–54. doi: 10.1109/CIASG.2011.5953331.
76. Haq R, Ni Z (2019) Classification of Electricity Load Profile Data and The Prediction of Load Demand Variability. *2019 IEEE International Conference on Electro Information Technology (EIT)*, 304–309.
77. Jeong HC, Kang BO (2019) Development of Characterization and Clustering Method of Daily Load Profiles for Time-of-Use (TOU) Tariff Structure. *2019 IEEE Power & Energy Society General Meeting (PESGM)*, 1–5. doi: 10.1109/PESGM40551.2019.8973623.
78. Azaza M, Fournier J, Lacarrière B, et al. (2017) Smart meter data clustering using consumption indicators : responsibility factor and consumption variability. *Energy Procedia* 142: 2236–2242. doi: 0.1016/j.egypro.2017.12.624.
79. Sun M, Konstantelos I, Strbac G (2017) C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data. *IEEE T Power Syst* 32: 2382–2393. doi: 10.1109/TPWRS.2016.2614366.

80. Ryu S, Choi H, Lee H, et al. (2019) Convolutional Autoencoder based Feature Extraction and Clustering for Customer Load Analysis. *IEEE Trans Power Syst* 35: 1048–1060. doi: 10.1109/TPWRS.2019.2936293.
81. Rajabi A, Eskandari M, Ghadi MJ, et al. (2019) A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications. *Energy Build* 203: 109455. doi: 10.1016/j.enbuild.2019.109455.
82. Rhodes JD, Cole WJ, Upshaw CR, et al. (2014) Clustering analysis of residential electricity demand profiles. *Appl Energy* 135: 461–471. doi: 10.1016/j.apenergy.2014.08.111.
83. Westermann P, Deb C, Schlueter A, et al. (2020) Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. *Appl Energy* 264: 114715. doi: 10.1016/j.apenergy.2020.114715.
84. Satre-meloy A, Diakonova M, Grünewald P (2020) Cluster analysis and prediction of residential peak demand profiles using occupant activity data. *Appl Energy* 260: 114246. doi: 10.1016/j.apenergy.2019.114246.
85. Jin L, Lee D, Sim A, et al. (2017) Comparison of clustering techniques for residential energy behavior using smart meter data. *AAAI Work. - Tech Rep*, 260–266.
86. Wang Y, Chen Q, Kang C, et al. (2017) Sparse and Redundant Representation-Based Smart Meter Data Compression and Pattern Extraction. *IEEE Trans Power Syst* 32: 2142–2151. doi: 10.1109/TPWRS.2016.2604389.
87. Savvopoulos A, Kalogeras G, Anagnostopoulos C, et al. (2020) Cluster-based Energy Load Profiling on Residential Smart Buildings. *25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 821–828. doi: 10.1109/ETFA46521.2020.9212161.
88. Wen L, Zhou K, Yang S, et al. (2018) Compression of smart meter big data: A survey. *Renew Sustain Energy Rev* 91: 59–69. doi: 10.1016/j.rser.2018.03.088.
89. Räsänen T, Voukantsis D, Niska H, et al. (2010) Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Appl Energy* 87: 3538–3545. doi: 10.1016/j.apenergy.2010.05.015.
90. Semeraro L, Giunta G (2014) Electrical Load Clustering : the Italian case. *IEEE PES Innovative Smart Grid Technologies, Europe*, 1–6. doi: 10.1109/ISGTEurope.2014.7028919.
91. Wang Y, Chen Q, Gan D, et al. (2018) Deep Learning-Based Socio-demographic Information Identification from Smart Meter Data. *IEEE Trans Smart Grid* 10: 2593–2602. doi: 10.1109/TSG.2018.2805723.
92. Choksi KA, Jain S, Pindoriya NM (2020) Feature based clustering technique for investigation of domestic load profiles and probabilistic variation assessment: Smart meter dataset. *Sustain Energy Grids* 22: 100346. doi: 10.1016/j.segan.2020.100346.
93. Ozawa A, Furusato R, Yoshida Y (2016) Determining the relationship between a household ' s lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles. *Energy Build* 119: 200–210. doi: 10.1016/j.enbuild.2016.03.047.
94. Panapakidis IP, Papadopoulos TA, Christoforidis GC, et al. (2014) Pattern recognition algorithms for electricity load curve analysis of buildings. *Energy Build* 73: 137–145. doi: 10.1016/j.enbuild.2014.01.002.
95. Motlagh O, Berry A, Neil LO (2019) Clustering of residential electricity customers using load time series. *Appl Energy* 237: 11–24. doi: 10.1016/j.apenergy.2018.12.063.

96. Freire VA, De Arruda LVR (2016) Identification of Residential Load patterns based on neural networks and PCA. *12th IEEE International Conference on Industry Applications (INDUSCON),* 1–6. doi: 10.1109/INDUSCON.2016.7874495.
97. Drechny M (2018) The method of consumers identification based on compressed power load profiles. *2018 Innovative Materials and Technologies in Electrical Engineering, i-MITEL 2018,* 1–4. doi: 10.1109/IMITEL.2018.8370464.
98. Cai X, Wang Y, Zhang J, et al. (2019) Data-Driven Load Data Cleaning and Its Impacts on Forecasting Performance. *iSPEC 2019 - 2019 IEEE Sustain Power Energy Conf Grid Mod Energy Revolution,* 1755–1760. doi: 10.1109/iSPEC48194.2019.8975047.
99. Capozzoli A, Savino M, Brandi S (2017) Mining typical load profiles in buildings to support energy management in the smart city context. *Energy Procedia* 134: 865–874. doi: 10.1016/j.egypro.2017.09.545.
100. Wang S, Chen H, Wu L, et al. (2020) A novel smart meter data compression method via stacked convolutional sparse auto-encoder. *Int J Electr Power Energy Syst* 118: 105761. doi: 10.1016/j.ijepes.2019.105761.
101. Selvam MM, Gnanadass R, Padhy NP (2017) Fuzzy based clustering of smart meter data using real power and THD patterns. *Energy Procedia* 117: 401–408. doi: 10.1016/j.egypro.2017.05.158.
102. Gavrilas M, Gavrilas G (2010) Application of Honey Bee Mating Optimization Algorithm to Load Profile Clustering. *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, 113–118. doi: 10.1109/CIMSA.2010.5611759.
103. Kim YI, Kang SJ, Ko JM, et al. (2011) A Study for Clustering Method to generate Typical Load Profile for Smart Grid. *8th International Conference on Power Electronics - ECCE Asia,* 1102–1109.
104. Rajabi A, Li L, Zhang J, et al. (2017) A Review on Clustering of Residential Electricity Customers and Its Applications. *20th International Conference on Electrical Machines and Systems (ICEMS),* 1–6. doi: 10.1109/ICEMS.2017.8056062.
105. Qiu W, Zhai F, Bao Z, et al. (2016) Clustering Approach and Characteristic Indices for Load Profiles of Customers Using Data from AMI. *China International Conference on Electricity Distribution,* 10–13.
106. Muthamizh M (2017) Fuzzy based clustering of smart meter data using power and THD Fuzzy based based clustering clustering of smart smart meter data using using real power and THD meter real power clustering of smart meter data using real power of smart meter data using real. *Energy Procedia* 117: 401–408. doi: 10.1016/j.egypro.2017.05.158.
107. Sarikprueck P, Attaphong C, Lumyong P, et al. (2017) Analyzing technique for electrical energy monitoring system in Thailand hospital. *Conf Proc - 2017 17th IEEE Int Conf Environ Electr Eng 2017 1st IEEE Ind Commer Power Syst Eur EEEIC / I CPS Eur,* 7–10. doi: 10.1109/EEEIC.2017.7977465.
108. Al-wakeel A, Wu J, Jenkins N (2017) k -means based load estimation of domestic smart meter measurements q. *Appl Energy* 194: 333–342. doi: 10.1016/j.apenergy.2016.06.046.
109. Xiang Y, Hong J, Yang Z, et al. (2020) Slope-based Shape Cluster Method for Smart Metering Load Profiles. *IEEE Trans Smart Grid* 11: 1809–1811. doi: 10.1109/TSG.2020.2965801.
110. Zarabie AK, Member S, Lashkarbolooki S, et al. (2019) Load Profile Based Electricity Consumer Clustering Using Affinity Propagation. *IEEE International Conference on Electro Information Technology (EIT),* 474–478.

111. Labeeuw W, Member GS, Deconinck G, et al. (2013) Residential Electrical Load Model based on Mixture Model Clustering and Markov Models. *2013 IEEE*, 1–9.

112. Oprea SV, Bara A, Lungu I (2015) Methods for electricity load profile calculation within deregulated markets. *2015 19th Int Conf Syst Theory, Control Comput. ICSTCC 2015 - Jt. Conf. SINTES 19, SACCS 15, SIMSIS 19*, 848–853. doi: 10.1109/ICSTCC.2015.7321400.

113. Mutanen A, Niska H, Jarventausta P (2016) Mining Smart Meter Data - Case : Finland. *CIRED*, 1–4.

114. Al-wakeel A, Wu J (2016) K-means based cluster analysis of residential smart meter measurements. *Energy Procedia* 88: 754–760. doi: 10.1016/j.egypro.2016.06.066.

115. Mamchych T, Wallin F (2014) Looking for patterns in residential electricity consumption. *Energy Procedia* 61: 1768–1771. doi: 10.1016/j.egypro.2014.12.208.

116. Nystrup P, Madsen H, Blomgren EMV, et al. (2021) Clustering commercial and industrial load patterns for long-term energy planning. *Smart Energy* 2: 100010. doi: 10.1016/j.segy.2021.100010.

117. Gouveia JP, Seixas J (2016) Unraveling electricity consumption profiles in households through clusters : Combining smart meters and door-to-door surveys. *Energy Build* 116: 666–676. doi: 10.1016/j.enbuild.2016.01.043.

118. Desgraupes B (2013) Clustering Indices. *CRAN Packag*, Available from: https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf.

119. Panapakidis IP, Christoforidis GC (2018) Optimal selection of clustering algorithm via Multi-Criteria Decision Analysis (MCDA) for load profiling applications. *Appl Sci* 8: 1–43. doi: 10.3390/app8020237.

120. Czétány L, Vamos V, Horvath M, et al. (2021) Development of electricity consumption profiles of residential buildings based on smart meter data clustering. *Energy Build* 252: 111376. doi: 10.1016/j.enbuild.2021.111376.

121. Steinley D, Brusco MJ (2011) Choosing the Number of Clusters in K-Means Clustering. *Psychol Methods* 16: 285–297. doi: 10.1037/a0023346.