

A comparative study on pretreatment methods and dimensionality reduction techniques for energy data disaggregation in home appliances

Viktor Isanbaev^a, Raúl Baños^{*a}, Francisco M. Arrabal-Campos^a, Consolación Gil^b, Francisco G. Montoya^a, A. Alcayde^a

^aDepartment of Engineering, University of Almería, Carretera de Sacramento, s/n, 04120 Almería (Spain)

^bDepartment of Informatics, University of Almería, Carretera de Sacramento, s/n, 04120 Almería (Spain)

Abstract

Energy meters provide valuable information that can be used to determine important features such as energy consumption of electrical devices and consumption habits in corporate, residential or public institutions. This information is crucial to establish energy saving strategies. With this aim, different approaches have been proposed in the literature, including non-intrusive load monitoring techniques, which enable the energy disaggregation of appliances and devices through a centralized measurement taken at panel level using a metering infrastructure. Generally, the accuracy of these techniques increases as more information is available on the analyzed signals or through subsequent post-computed values. Active power, reactive power, or even current harmonics measurements can be used for this task. However, the use of these and other recently proposed power and current features increases the dimensionality and, therefore, the complexity of the algorithms involved in the disaggregation process. Therefore, it is necessary to apply advanced techniques to reduce the dimensionality of the data, as well as the possible linear dependence between variables. This paper compares the performance of 8 data pretreatment methods and 6 dimensionality reduction techniques to data retrieved by an advanced metering infrastructure in a real environment consisting of 10 different home appliances. Results obtained from the comparative analysis show that the information provided by raw data can be enhanced by using pretreatment techniques and dimensionality reduction methods, especially when a custom combination of active power and current harmonics measures is considered.

Keywords: Energy monitoring, energy data disaggregation, pretreatment methods, dimensionality reduction, harmonics, home appliances.

*Corresponding author

Email addresses: vs613@ual.es (Viktor Isanbaev), rbanos@ual.es (Raúl Baños*), fmarrabal@ual.es (Francisco M. Arrabal-Campos), cgilm@ual.es (Consolación Gil), pagilm@ual.es (Francisco G. Montoya), aalcayde@ual.es (A. Alcayde)

1. Introduction

The efficient use of energy is one of the most crucial and challenging issues according to the UN Sustainable Development Goals (SDG) [1, 2]. Energy efficiency strategies are focused from generation and consumption viewpoints. Thus, renewable energy systems are being incorporated into the grid to reduce dependency on fossil fuels, while energy saving strategies are being promoted to reduce the consumption of energy and, therefore, the volume of electricity that should be generated. The synergy of these actions will have a global impact in terms of reduction of emission of polluting gases. Recent studies have highlighted the possibilities of Internet of Things (IoT) and advanced big data analytics for improving energy efficiency of smart home appliances [3]. Energy big data is characterized by the need of processing large volumes of continuous stream of data from a large variety of sources [4]. Some challenges in the field of big data in microgrids have also been identified, including the need of developing advanced sensors technologies that allow to operate with the streaming data, efficient techniques for data storage, data security, data visualization, etc. [5].

As far as energy efficiency strategies are concerned, their application would be reflected in a reduction of the monthly energy bill for businesses, public institutions and families. For example, energy efficiency can be achieved by using electrical appliances in certain (reduced-price) time slots or deactivating appliances when they are not needed. Furthermore, unplugging devices when not in use is an important issue, which is why some researchers have analyzed how to determine occupancy profiles on basis to the appliance usage [6, 7]. But these and other energy saving strategies require considering efficient procedures for monitoring electrical appliances. The most evident way to perform this task is to implement Intrusive Load Monitoring (ILM) techniques [8], which are based on the use of low-end electricity meter devices directly measuring each appliance or device. An alternative option is to apply Non-Intrusive Load Monitoring (NILM) techniques [9], sometimes referred to as load disaggregation, which is attracting an increasing interest due to the roll-out of metering technology around the world [10, 11]. The aim of NILM techniques is to identify the different loads connected simultaneously to a power source, estimate the current status and report a detailed disaggregated power consumption using a single energy meter installed in the main electric panel. But in addition to applying NILM for energy saving purposes, the activity monitoring applications through NILM is recently receiving much interest [11]. The main differences between the two techniques lie in the cost of the solution, its deployment, and maintenance. In fact, ILM adds complexity to the acquisition stage and is more expensive, while NILM method is more scalable at the cost of a higher computational effort and the need of a high-precision measuring device.

An important barrier to apply ILM and NILM at large scale is the high-price of commercial energy meters. This is the reason why some researchers are proposing low-cost and high-precision energy meters [12]. In addition to the large volume of data, the application of NILM techniques requires the development of new methods that can effectively decrease the additional dimensionality introduced by the new variables and, therefore, to reduce the computation time and the possibility of overfitting. Visualizing high dimensional data is a common problem between many research domains, such as hyperspectral imaging [13] or audio unmixing [14]. Dimensionality Reduction (DR) techniques [15] allows to transform data from a high-dimensional space into a low-dimensional space so that the latter retains some relevant properties of the original data. This characteristic is used in our investigation to explore the separability between appliances and, particularly, their data structure. It should be noted that dimensionality reduction methods have been successfully applied in the past to analyze electricity consumption profiles [16, 17].

The aim of our investigation is to determine the advantages and disadvantages of different linear and nonlinear dimensionality reduction (DR) methods and pretreatment techniques to manage three electrical harmonic feature sets (current harmonics, power harmonics and a customized combination of them) considering real data of home appliances. The contributions of this paper can be summarized as follows:

- It proposes the use of complex harmonic features as input variables for NILM methods. The use of harmonics for NILM purposes is a promising strategy that is attracting attention in recent years [11, 18, 19, 20, 21]. In [20], the researchers use harmonic features, in particular the 3rd, 7th and 11th current harmonics in combination with Principal Components Analysis (PCA). In our study, however, the current and voltage signals sampled at high frequency are computed to obtain the harmonic characteristics of both up to order 50. In addition, power harmonics and the $\cos(\varphi)$ up to order 50 are also obtained. For this purpose, the FFT algorithm implemented in the metering infrastructure (openZmeter, oZm [22]) has been used. Through the built-in API, it is possible to get the data in order to obtain the magnitude and phase angle for each of the complex harmonics. In addition, the active power levels of each harmonic have also been measured, thus providing another subset of measurements related to the harmonic distribution of the signal. Finally, a comparison between two different harmonic feature sets and a customized combination of them is presented.
- It includes a comparative analysis of the effect of data pretreatment techniques applied to feature sets. Eight data pretreatment techniques (Centering, Autoscaling, Range scaling, Pareto scaling, Vast scaling, Level scaling, Log transformation and Power transformation) often used in previous investigations [23] are considered in our study. The aim is to find the most appropriate technique

to obtain the best regularization results in the data fed to the subsequent dimensional reduction algorithms.

- It includes a comparative analysis of the performance of dimensionality reduction (DR) methods. The pretreated data is fed to six linear and nonlinear DR approaches (Principal Components Analysis, Linear Discriminant Analysis, Fast Independent Component Analysis, Partial Least Squares Regression, t-Distributed Stochastic Neighbor and Uniform Manifold Approximation and Projection). It is worth mentioning that the authors of [20] propose, as future work, to compare PCA with comparison of other dimension reduction algorithms such Linear Discriminant Analysis (LDA), aspect that is covered in the present study.
- It includes an empirical study in a real-life environment involving a wide range of modern appliances, from basic loads to time/state adjustable loads. Furthermore, it is shown that the analysed methods can also be successfully applied to existing data in other public databases.

The rest of the paper is organized as follows. Section 2 gives an overview of the latest advances on engineering optimization algorithms related to electricity consumption disaggregation of appliances. Section 3 describes the materials that have been used to acquire the harmonic feature sets from different home appliances and how pretreatment techniques and DR methods have been computed. Section 4 presents the results obtained in the experimental study and a detailed comparison of numerical and graphical results of the performance of these pretreatment methods and DR techniques applied to a set of household appliances. Finally, Section 5 presents a discussion about the implications and findings for NILM, as well as the main conclusions, limitations and future investigations derived from this study.

2. Related work

Engineering applications of NILM have attracted the attention of many researchers in the last years [24]. In particular, there has been a surge in interest in developing novel computational approaches in the field of sensors technologies focused to NILM [25]. Device-level consumption footprints inferred by NILM can provide statistics with personalized consumption of each appliance through exploiting the aggregated power signal and help to promote energy saving actions [26]. NILM [9, 27] is often divided into four phases [28]:

- **Data acquisition:** In this phase, the electrical signal is analyzed in order to calculate the electrical quantities. This requires installing a metering device in the electrical panel of a home or business to collect data on voltage, current, active power, reactive power, power factor, etc.

- **Feature extraction:** The objective of this phase is to extract those features necessary for identifying the state (switched-on or switched-off) of the loads or the detection of events, such as state transitions of household appliances. Feature extraction often requires the analysis of time series [29, 30] and the application of dimensional reduction techniques to transform a large number of features to a lower dimensionality space [20, 31, 32].
- **Inference and learning:** Once the necessary characteristics have been extracted, the objective is to use them to identify the appliances that are currently operating. This requires to apply disaggregation algorithms in order to categorize the appliances that are currently in use.
- **Appliance classification and load disaggregation:** This phase, also known as load identification, aims to determine the operating state of the appliances. It often requires dividing the total consumption among the identified loads is required [28].

Some authors have analyzed in detail the challenges of NILM in practical applications [33]. On the one hand, it is required to use an accurate metering infrastructure with high sampling rates for data acquisition and feature extraction, capable of processing the aggregate signal containing data of all the appliances connected to the network [34]. In fact, some authors have highlighted that neither the reactive power nor the high-sampling measurement is the standard feature of the currently available smart meters [35]. An important contribution of this paper is to use a high-precision and reliable power and electric energy meter to acquire and process data.

On the other hand, the efficiency of non-intrusive appliance load identification methods [36, 37, 38] is highly dependent on the selected features. Almost all approaches found in the literature apply electricity consumption disaggregation, such that the feature space is partitioned according to the appliance level of active or reactive power consumption from aggregate measurements of voltage and current obtained at a centralized location in the electrical panels [39], in some cases adding individual devices in the appliances (for example, in [40] it is used a electromagnetic field (EMF) sensor to measure the magnetic and the electrical fields nearby each appliance in order to detect its operational state). A typical approach is to apply optimization or pattern recognition-based approaches [41] to determine the best combination of the appliance load to estimate the total power consumption [42]. Despite the above, the disaggregation of energy consumption to identify loads presents some problems, particularly the difficulty of differentiating between devices that have a similar level of consumption. Other alternatives to power-based frameworks are voltage and current-based criteria [43]. However, an open research question is to determine whether or not other

power quality variables could be successfully used for this purpose with real home appliances. A few studies have determined that harmonics can be useful to track the power consumption of the variable load and, in turn, to disaggregate this variable load from the power consumption [44]. An important contribution of this paper is the application of FFT to current and voltage signals sampled at high frequency and the use of the features derived from the harmonic characteristics obtained. In particular, current harmonics and active power harmonics of the two signals are considered.

Finally, it is also important to remark that a significant number of synthetic or real datasets have been proposed in the past for NILM purposes. For example, the Reference Energy Disaggregation Data Set (REDD) [45] is a public data set containing detailed power consumption from six real homes, for the whole house as well as for each individual circuit in the house. Other repository is UK-DALE [46], an open-access data set that provides the power demand from five houses, such that the both, the whole-house mains and individual appliances power demand, are recorded every six seconds. Other open-access dataset is provided by the Indian Dataset for Ambient Water and Energy (iAWE) [47], which contains the aggregated and sub-metered electricity measurements (one-second resolution) retrieved from a house during more than two months. Additional information about datasets used for electricity consumption disaggregation can be found in [48, 49, 33]. However, only a few datasets provide pre-computed harmonic content. One example is EMBED dataset [50]. EMBED provides data in their aggregated form, which do not allow to use them for studying data pre-processing techniques and defining the footprint of an individual appliance. EMBED offers plug load data, but the 1-2Hz sampling frequency does not allow harmonic extraction according to the Nyquist theorem. Other datasets offer disaggregated data from individual devices recorded at high frequencies, which allows the harmonics to be extracted. In some datasets, different switch-on scenarios have been included. A drawback of some datasets is that the duration of the recordings is only a few seconds, and it is not possible to extract statistical data on the extended operation of the device.

3. Materials and methods

3.1. Materials

This study focuses on the analysis of a set of home appliances commonly used in households. A group of appliances with resistive features, such as the oven, oil heater, grill and toaster are considered, while other group of non-resistive appliances with inductive characteristics are also considered, including vacuum cleaner and kitchen hood, as well as other whose main load is a compressor (electric motor load) such as freezers and refrigerators. Finally, two devices that rely on power electronics (a television and a laptop charger) are

also analyzed. The full list of appliances with brand, model, nominal active power and the tag used within the code can be found in Table 1.

Some authors have revised the literature of machine learning methods applied to smart-building applications [51], where energy devices (including appliances and sensors) are analyzed from different perspectives, including: (i) the energy profiling and demand estimation; and (ii) appliances profiling and fault detection.

Appliance	Tag	Brand	Model	Power (W)
Kitchen hood	kitchen_hood	Teka	c-620	350
Vacuum cleaner	vacuum_cleaner	Tristar	SZ-2174	800
Laptop charger	laptop	Lenovo	ADL65YCC3A	65
Television	tv	Toshiba	26AV615DG	120
Refrigerator	fridge	Brandt	BFD6425BW	75
Freezer	freezer	Hisense	FT124D4HW1	168
Oven	oven	Zanussi	ZOB20311XU	1875
Grill	grill	Corberó	CPA-4700	2000
Toaster	toaster	Almison	ALMTOS25C	2400
Oil radiator	oil_heater	Haeger	Zen XI OH-011.002A	2500

Table 1: List of domestic appliances used for the experiments.

The harmonic feature sets related to the operation of the home appliances is retrieved using the *openZmeter* (oZm) [22]. oZm is an open-source and open-hardware single-phase electrical energy meter and power quality analyzer with IoT capabilities that allows measuring a wide range of electrical variables such as RMS voltage and current, active, reactive and apparent power, current and voltage harmonics up to order 50, Total Harmonic Distortion (THD), power factor, etc. The data retrieved by oZm can be accessed through a simple web interface or through a specific Application Programming Interface (API). It is capable of sampling the voltage and current waveform at a frequency of up to 15625Hz and compute parameters according to IEC61000-4-30 and EN-50160 regulations. It is noticed that oZm collected independent measurements for each home appliance during 10 minutes. Different time span aggregations of 200 ms, 3 seconds, 1 minute, 10 minutes, 15 minutes and 1 hour are generated automatically and gathered by oZm (200ms is the interval used in this study). Python [52] and Scikit-learn [53] (along with many other libraries) are used to perform data acquisition, pretreatment, application of dimensionality reduction methods and finally result evaluation of the available data.

3.2. Methods

Overall workflow of the data throughout the experiment can be observed in Figure 1. The process starts with the data acquisition stage using the oZm meter where mainly harmonic content is gathered

and pre-processed using custom developed scripts. Additionally, the measurements are then applied to the pretreatment functions. Finally, several dimensionality reduction methods (2D) are applied afterwards to compare the results using the evaluation metrics.

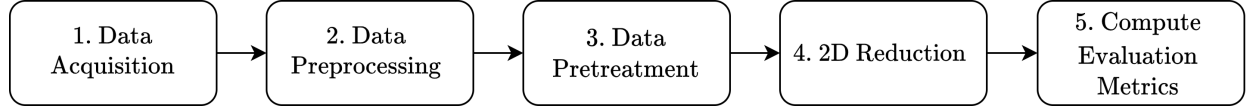


Figure 1: Experiment data workflow summary. Data passes through five stages, starting with the acquisition, followed by its preprocessing and pretreatment, to then apply DR methods and evaluate the result.

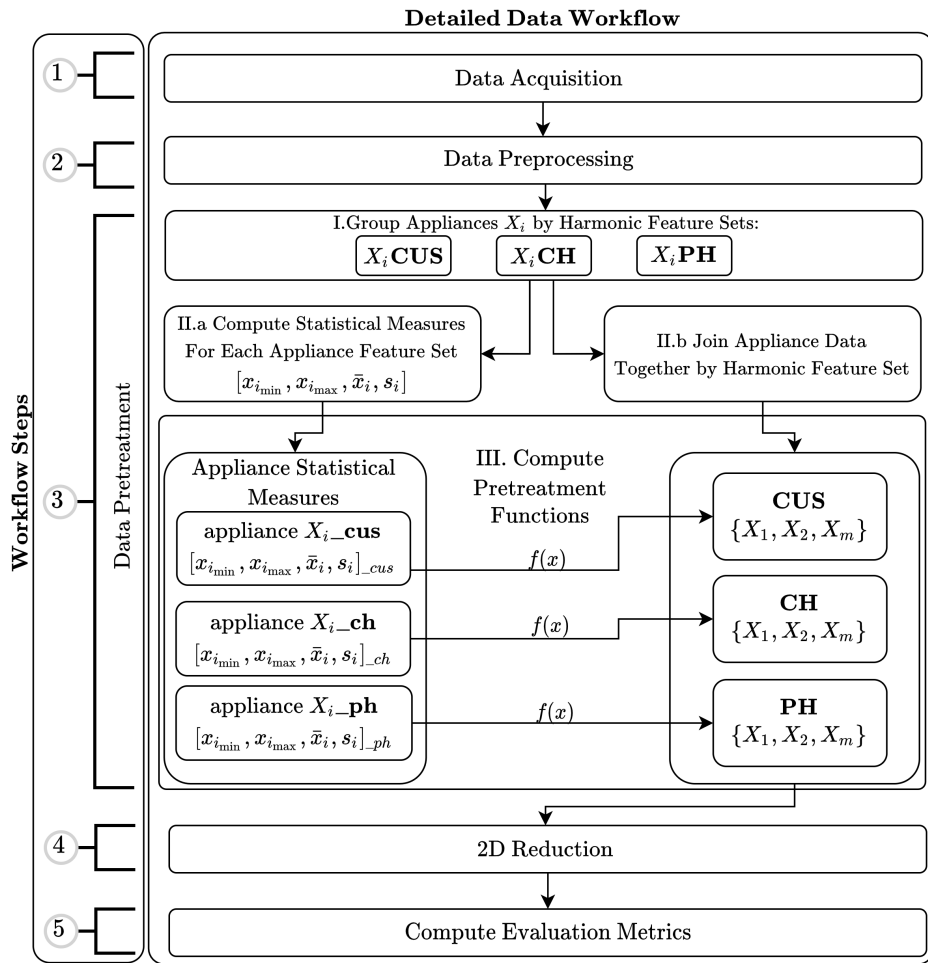


Figure 2: Detailed description of the data pretreatment.

As described in Figure 2, the data is acquired and preprocessed for each appliance listed in Table 1 and then passes to the data pretreatment stage. By aggregating the data of electrical appliances, three datasets of harmonic features (*ch*, *ph*, *cus*) are created in different steps: I) For each appliance, three groups of harmonic

features (see Table 2) are obtained $[x_i\text{ch}, x_i\text{ph}, x_i\text{cus}]$; II.a) Statistical measures for each appliance harmonic feature set are computed; II.b) The appliance harmonic features are appended according to harmonic feature type (ch , ph , cus); III) Pretreatment functions described in section 3.2.3 are applied to each dataset. Once the data has been preprocessed and pretreated, dimensional reduction methods described in Section 3.2.4 are applied to the datasets. Finally, after applying dimensionality reduction methods, the clustering evaluation metric *Silhouette Score* [54] is applied. The next subsections will go over each of the stages in further detail.

3.2.1. Data acquisition

Figure 3 shows the measurement scheme. As it is shown, a single oZm device measures the energy and power quality variables. The information acquired and processed by oZm is submitted via Wi-Fi to a computer, which is in charge of applying preprocessing, pretreatment, dimensionality reduction methods to these data. In this experiment, independent measurements are applied to the 10 home appliances, i.e., when one appliance is measured, the rest are switched off.

The model presented here makes use of data from recording several minutes of operation of the appliances listed in Table 1 using the oZm API, which collects complex harmonic values for voltage and current up to order 50. Most papers found in the literature consider active and reactive power as features to model load behavior [55]. Some authors have also proposed the use of distortion power in addition to the active and reactive power to obtain models of appliance classes using dimensionality reduction techniques [11, 40]. Nonetheless, data related to the total active or reactive power is not considered here since one of the aims of the paper is to show that harmonics can provide sufficient information to reach our goals. The data for each electrical appliance shown in Table 1 is gathered from the oZm device in a manner that each i -th appliance has a separate dataset of n measurements, $X_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$. Then, the data is filtered by applying a fixed threshold value to the total RMS current to discard the idle state of these appliances. Afterwards, for each appliance dataset X_i , three sets of harmonic features are extracted for the experiment, i.e., current harmonics (ch), active power harmonics (ph) and a custom combination (cus) of both features.

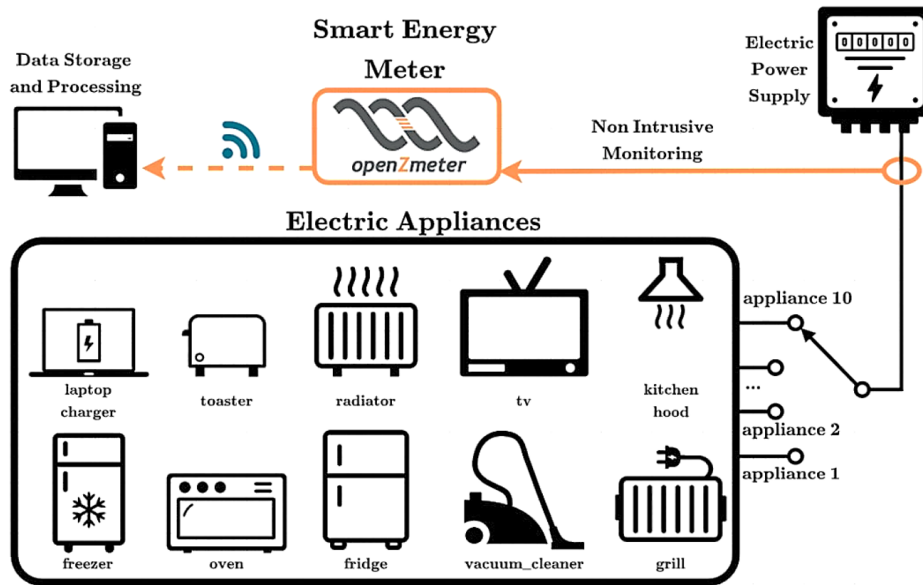


Figure 3: Physical measurement scheme.

In order to obtain a set of structured data for a given time interval, a post request is performed to the oZm API. The request includes the time interval in timestamp format with millisecond precision, identification number of the analyzer, cookie and the list of features to be analyzed, as in the following example:

```
curl --location --request POST 'http://192.168.5.51/getSeries' \
--header 'Accept: application/json' \
--header 'Content-Type: application/json' \
--header 'Cookie: SESSION=8FE265C5-8BB8-47FD-B7E4-12F6AE9858AA' \
--data-raw '{ "Aggreg": "200MS",
  "Analyzer": "F10012009034E42413032302_CH1_1P",
  "From": 1620057540000,
  "To": 1620057600000,
  "Series": [ "Frequency", "Unbalance",
    "Flag", "Voltage", "Current", "Active_Power",
    "Reactive_Power", "Power_Factor", "Phi",
    "Voltage_Harmonics_Complex",
    "Current_Harmonics_Complex",
    "Power_Harmonics" ] }'
```

Listing 1: oZm post request example to retrieve 60s of data in aggregations of 200ms.

After this request execution, oZm returns a response in a JSON array form. The first position in the array contains the header of the subsequent entries. As it can be observed below, the first position contains

the timestamp for the current entry:

```
[
  ["Time", "Frequency", "Unbalance", "Flag", "Voltage", "Current", "Active_Power",
   "Reactive_Power", "Power_Factor", "Phi",
   "Voltage_Harmonics_Complex",
   "Current_Harmonics_Complex",
   "Power_Harmonics"],
  [1620057540076, 50.0438, 0, false, [242.478], [0.017807], [-0.094611],
   [1.92934], [-0.021912], [-92.8224],
   [[-3.72777,242.393], [0.004717,0.043773], [-1.93943,-0.325571], [...]],
   [[0.007964,-0.00027], [-9.2e-05,0.000262],[0.000231,0.000135], [...]],
   [-0.095116, 1.1e-05, -0.000491, [...]]
], [...], [...]]
```

Listing 2: oZm returns a response in a JSON array form. The first element is the data header and the following subarrays contain the data sorted according to the header.

3.2.2. Data preprocessing

The raw data obtained requires some preprocessing before applying pretreatment methods. At a first glance, the preprocessing and pretreatment process may seem similar as both prepare the data to apply further actions. However, preprocessing handles the extraction of harmonic features, filtering and preparing the datasets that will be used further for the experiment, while pretreatment only applies the data enhancement functions.

During the preprocessing stage, the following tasks were carried out:

1. **Idle threshold.** Electric devices have idle states when the main component is turned off. It occurs, for example, when a certain setpoint of the temperature of the oven is established. To filter out this kind of data, a fixed threshold of 0.2A is set to consider only current measurements above this value.
2. **Data Homogeneity.** The number of records in each dataset can vary for each electrical appliance depending on the elapsed measurement time span. To avoid conflicts in the data management, a fixed number of records is established for all datasets in order to create homogeneous 2D plots.
3. **Harmonic Preprocessing.** oZm computes complex harmonic values in Cartesian's real and imaginary form using the FFT algorithm, that is, $\mathbf{i}_k = i_{ak} + ji_{rk}$, where i_{ak} denotes the real part and i_{rk} the imaginary part of a complex number \mathbf{i}_k for the k -th harmonic. However, since the management of

absolute values and phase angles is more suitable, the calculation used instead is:

$$i_k = \sqrt{i_{ak}^2 + i_{rk}^2} ; \varphi_k = \arctan \frac{i_{rk}}{i_{ak}} \quad (1)$$

One of the goals of this paper is to compare the different combinations of harmonic features. To achieve this task, three different groups of harmonic features are considered (see Table 2):

- **Current Harmonics (CH):** Harmonics that represent the distortion of the current due to nonlinear loads. The API of oZm provides the data with a timestamp and an array of 50 complex values, each containing the real and the imaginary part.
- **Active Power Harmonics (PH):** The active power of each harmonic computed by the oZm. It is obtained in array form, but contrary to the current harmonics, this array is composed exclusively of real numbers using the expression $P_k = V_k I_k \cos \varphi_k$ where P_k is the harmonic active power, V_k the harmonic RMS voltage, I_k the harmonic RMS current and φ_k is the phase angle between voltage and current of the k -th harmonic.
- **Custom Harmonics Combination (CUS)** A combination of three significant electric features related to the k -th harmonic order was computed, resulting in an array of a total of 150 variables. The three features $[I_k, \varphi_k, P_k]$ are the current harmonic, phase angle between the current and voltage waveforms and the active power harmonic of order k , respectively.

Type of Data	Tag	Description
Current harmonics	<i>ch</i>	Value of the first 50 current harmonics
Power harmonics	<i>ph</i>	Active power for the first 50 power harmonics
Custom harmonics combination	<i>cus</i>	Combination of first 50 active power and current harmonics

Table 2: Combinations of features used in the model (the first 50 harmonics have been considered).

Linear loads are those that have a linear relationship between voltage and current. This means that current and voltage share the same harmonic content for a given time span. Electric heaters and incandescent light bulbs are an example of such loads. They normally produce a current waveform that mimics the spectral content of the voltage. Under ideal conditions (perfect sinusoidal voltage), the current does not include harmonic content. On the contrary, nonlinear loads change their impedance with time which means that the generated current does not have a sinusoidal form. This implies the presence of harmonic components in the current and, depending on the grid, also on the voltage

supply waveform. Examples of nonlinear loads are laptops or smartphone chargers, which contain voltage rectifiers to transform AC to DC voltage.

Although harmonics are the cause of power quality problems in the power grid, they also contain rich information about the electrical signature of the appliances. In this sense, the oZm applies a Fourier Transform to the voltage and current time series in order to analyze them in the frequency domain.

3.2.3. Data pretreatment

An important issue of this investigation is to compare different pretreatment functions to the data acquired by oZm. This analysis requires involves a series of stages:

1. **Group appliance by harmonic features:** Extension of Harmonic Preprocessing step to highlight that statistical measures will be computed over individual appliance feature sets (three feature sets for each appliance).
2. **Compute statistical measures:** Several statistical measures (minimum, maximum, mean and standard deviation) are obtained from the pretreatment functions described in Table 2. These statistical measures are computed for each appliance feature set separately, such that one statistical measure set $[x_{i_{\min}}, x_{i_{\max}}, \bar{x}_i, s_i]$ is obtained for each appliance feature set, and one additional set is generated for the combined dataset.
3. **Join appliance data by feature set:** From this point, data of individual appliances will no longer be required, but rather their aggregated measure. Therefore, three datasets with the features from Table 2, "cus", "ch" and "ph" will be considered.
4. **Compute pretreatment functions:** Centering, scaling and transformation functions [23, 56] are a set of pretreatment techniques [57] that use statistical and multi-variant methods for information extraction and data interpretation. Besides the application of the pretreatment functions, the mean (\bar{x}_i), standard deviation (s_i), minimum and maximum are transformed, such that the statistical measures of one appliance are applied to the entire features dataset composed of combined appliance harmonic features. The transformation process of the statistical measures consists in taking each appliance (one by one) from the combined dataset and to apply the pretreatment function using its $[x_{i_{\min}}, x_{i_{\max}}, \bar{x}_i, s_i]$ values. Table 3 shows the transformation (pretreatment) functions considered in this study, where \tilde{x} and \hat{x} are the data after applying these pretreatment steps:

Method	Tag	Formula
Raw	raw	x_{ij}
Centering	scale_center	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$
Autoscaling	scale_auto	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$
Range scaling	scale_range / scale_range_ms	$\tilde{x}_{ij} = \frac{s_i (x_{ij} - \bar{x}_i)}{(x_{i_{\max}} - x_{i_{\min}})}$
Pareto scaling	scale_pareto	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$
Vast scaling	scale_vast	$\tilde{x}_{ij} = \frac{(x_{ij} - \bar{x}_i)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$
Level scaling	scale_level	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$
Log transformation	scale_log	$\hat{x}_{ij} = \log_{10}(x_{ij})$
Power transformation	scale_power	$\hat{x}_{ij} = \sqrt{x_{ij}}$
		$\tilde{x}_{ij} = \hat{x}_{ij} - \bar{\hat{x}}_i$

Table 3: Transformation functions based on statistical parameters mean (\bar{x}_i), standard deviation (s_i), min ($x_{i_{\min}}$) and max ($x_{i_{\max}}$). The original value x_{ij} (raw data) is transformed to \tilde{x}_{ij} . Variable x can refer to current amplitude I , phase angle φ or active harmonic power P_i .

- (i) **Raw** data serve as a reference to compare results and determine if pretreatment has resulted in any improvement.
- (ii) **Centering** strategy consists of subtracting the mean value after which the data is centered at the origin. It is focused on differences.
- (iii) **Autoscaling** strategy is an extension of the previous method that additionally normalizes the standard deviation of data to 1. It compares harmonics based on correlation, although as a side effect, it produces inflation of measurement errors.
- (iv) **Range scaling** strategy divides the values by the difference between the maximum and minimum value. These additional range parameters, min and max, can also be substituted. Therefore, two variants of range scaling function can be considered: "scale_range", in which $[x_{i_{\min}}, x_{i_{\max}}, \bar{x}_i, s_i]$ are substituted, and "scale_range_ms" which substitutes $[\bar{x}_i, s_i]$ and uses the range values $[x_{i_{\min}}, x_{i_{\max}}]$ of the x_{ij} -th appliance.
- (v) **Pareto scaling** strategy is similar to autoscaling, and reduces the relative importance of large values, while the data structure is partially preserved. It is sensitive to large harmonic value changes.
- (vi) **VAST scaling** strategy is based on the Variable Stability procedure [58], which is focused to harmonics with little fluctuations.
- (vii) **Level scaling** strategy is similar to centering but it divides the values by the average value. It is focused on relative response and increases measurement errors.

- (viii) **Log transformation** strategy, widely used to deal with skewed data [59], subtracts the mean value from the data, similarly to the centering function to focus on the differences. Data for these two function required mapping. The values of the variables related to the angle between voltage and current of the CUS dataset were not mapped and the transformation was not applied due to the homogeneity of the values which are well defined in the interval $[-\pi, \pi]$. Since this transformation suffers problems with zero and negative values, an affine transformation [60] was performed over the rest of CUS features, in which values were mapped in a range from $[min_{all_values}, 0]$ to $[0.01 * min_{positive_values}, min_{positive_values}]$.
- (ix) **Power transformation** strategy has a similar effect to that of scaling-based transformation. It brings the scale closer together by reducing the differences between the dataset values, such that the large values become smaller in a larger scale compared to smaller values. As it presents similar problems to log transformation, same range transformation solution was applied.

3.2.4. Dimensionality reduction methods

Before applying dimensionality reduction techniques, it is important to remark that there are 50 variables for the datasets of current harmonic and power harmonic features, while there are 150 variables for the custom combination of harmonic features. In the pursuit to comprehend the behavior of appliance harmonics data and the pretreatment impact, the dimensionality is reduced to two dimensions (2D). Table 4 describes the DR methods used our investigation:

- **Principal Components Analysis (PCA)** [61] is a multivariate technique for reducing the dimensionality of large datasets that produces linear transformations of correlated variables by solving the eigenvector/eigenvalue problem. The resulting variables are called *principal components*. Some authors have used PCA as a preprocessing step before to apply machine learning methods [62], while in other cases PCA is used to reduce the dimension of power features [20].
- **Linear Discriminant Analysis (LDA)** [63] is a dimensionality reduction method which finds an optimal linear transformation that maximizes the class separability. The aim is to maximize the between-class scatter and minimize the within-class scatter. The number of data items must be higher than its original dimensionality.
- **Fast Independent Component Analysis (FICA)** [64] aims to find a linear representation of non-Gaussian data so that the components are statistically independent, or as independent as possible.

Such a representation seems to capture the essential structure of the data in many applications, including feature extraction and signal separation.

- **Partial Least Squares Regression (PLSR)** [65] combines some ideas of principal component analysis (PCA) and multiple linear regression strategies. Its objective is to derive a set of dependent variables from a collection of independent variables (predictors), such that this prediction is performed by extracting a collection of orthogonal components termed latent variables from the predictors that have the highest predictive ability.
- **t-Distributed Stochastic Neighbor (tSNE)** [66] visualizes high-dimensional data by giving each datapoint a location in a two- or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding [67] that produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map.
- **Uniform Manifold Approximation and Projection (UMAP)** [68] is a novel manifold learning technique for dimensionality reduction. UMAP is constructed from a theoretical framework based on Riemannian geometry and algebraic topology. The UMAP algorithm is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with higher runtime performance. Furthermore, UMAP has no computational restrictions on embedding dimensions, making it viable as a general-purpose dimension reduction technique for machine learning.

Dimensionality Reduction Method	Acronym	TAG
Principal Components Analysis	PCA	pca
Linear Discriminant Analysis	LDA	lda
Fast Independent Component Analysis	FICA	fica
Partial Least Squares Regression	PLSR	plsr
Uniform Manifold Approximation and Projection	UMAP	umap
t-Distributed Stochastic Neighbor	tSNE	tsne

Table 4: Dimensionality Reductions Methods and tag used in figures and cluster plots.

3.2.5. Evaluation metrics

In order to quantify and compare the pretreatment methods described above, a clustering validity metric will be applied to the results obtained from the dimensionality reduction methods. More specifically, *Silhouette Score* [54] is a normalized validity index which computes the result based on the intra-cluster and the inter-cluster distance, such that the score is bounded to a range $[-1, 1]$. Equation 2 describes how *Silhouette Score* is calculated, being a the intra-cluster distance (the mean distance to all other points within

the same cluster) and b the the inter-cluster distance (the mean distance to all other points in the nearest cluster). Scores equal or below 0 indicate overlapping clusters, being -1 the worst score and 1 the best one (non-overlapping). In fact, some studies have shown that this metric obtained the best results in a comparison involving 30 different Cluster Validity Indexes [69].

$$SS(i) = \frac{b - a}{\max(a, b)} \quad (2)$$

It is important to note that the original range of values that Silhouette Score can take are in the range $[-1,1]$. However, the initial score range $[-1, 1]$ is normalised to $[0, 100]$ in order to avoid negative numbers and to prevent representation difficulties that could arise in charts such as box plots. This is why the legend Silhouette Score is used with an asterisk (Silhouette Score*).

4. Results

This section presents the numerical and graphical results considering the methodology described above.

Numerical results are presented in table format, where rows (records) indicate the reduction method and the pretreatment function, while columns (attributes) indicate the feature set and the appliance name considered in each record. For example, Figure 4 presents the results obtained by LDA when it is applied in combination with the pretreatment methods (from "scale_center" to "scale_powertransform") considering current harmonics and using statistical measures according to the column header. In reference to the columns, those corresponding to home appliances (from kitchen_hood to oil_heater) indicate that the pretreatment function used is applied taking as reference the statistical values of the corresponding appliance of each column. Column labeled with "self" indicate that each appliance within a dataset considers its own statistical values during pretreatment and not a fixed values of one appliance, as in the case of columns from "kitchen_hood" to "oil_heater". The statistical measures of datasets labeled "global" are calculated for the whole dataset and not for each appliance.

method		Current Harmonics (CH)												
		$f(x)$: scale_	kitchen_hood	vacuum_cleaner	laptop	tv	fridge	freezer	oven	grill	toaster	oil_heater	global	self
LDA	raw	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89
	center	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	44.31	77.89
	auto	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	42.34	77.89
	range	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	38.13	77.89
	range_ms	90.31	97.50	96.56	83.01	83.10	78.42	86.21	87.97	85.13	89.67	38.13	94.42	
	pareto	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	45.59	77.89
	vast	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	40.28	77.89
	level	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	77.89	42.49	77.89
	logtransform	79.47	79.47	79.47	79.47	79.47	79.47	79.47	79.47	79.47	79.47	79.47	42.94	79.47
	powertransform	77.45	77.45	77.45	77.45	77.45	77.45	77.45	77.45	77.45	77.45	77.45	42.43	77.45

Figure 4: LDA 2D reduction method applied to CH features set to which pretreatment functions and substitution of statistical measures are applied.

Figure 5 shows all the numerical results obtained in our experiments. In order to facilitate the interpretation of the results, a custom heat map is used, in which only the scores greater than the raw results are highlighted. The orange color highlights the best global solution included in the results, green highlights the best value per column, and blue highlights scores that improve the score from "raw" row within 2D reduction method and feature set selection. The colors blue and green are based on a palette of 10 shades, with darker shades being used to represent higher values of Silhouette Score metric.

4.1. Impact of pretreatment functions on 2D dimensionality reduction methods

The application of pretreatment functions described in Table 3 yield better results compared to raw data, as it can be observed in the box and whiskers diagram shown in Figure 6, where only the functions outperforming the raw data results are displayed. The "range_ms" version of the Scale Range function is the best for linear DR methods, while "logtransform" performs well with nonlinear UMAP and tSNE.

Figure 7 is a modification of Figure 6 that shows the mean values for each pretreatment function and dimensionality reduction method. Overall, linear methods (LDA, PCA, FICA and PLSR) outperform the results obtained by non-linear approaches (UMAP and tSNE). Figure 8 shows the average effect of the pretreatment function in the interval [0,100], where 70.2 is the average score obtained for all the methods. It is clear that the "range_ms" version of the Scale Range function outperforms to the other pretreatment functions.

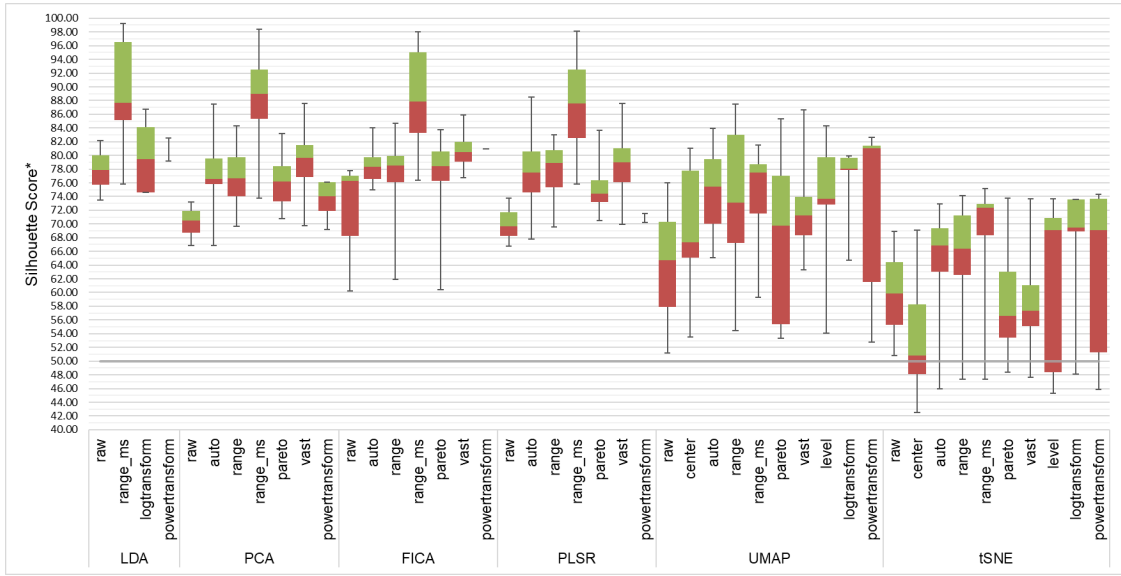


Figure 6: Silhouette Score* obtained using different pretreatment functions and dimensionality reduction methods.

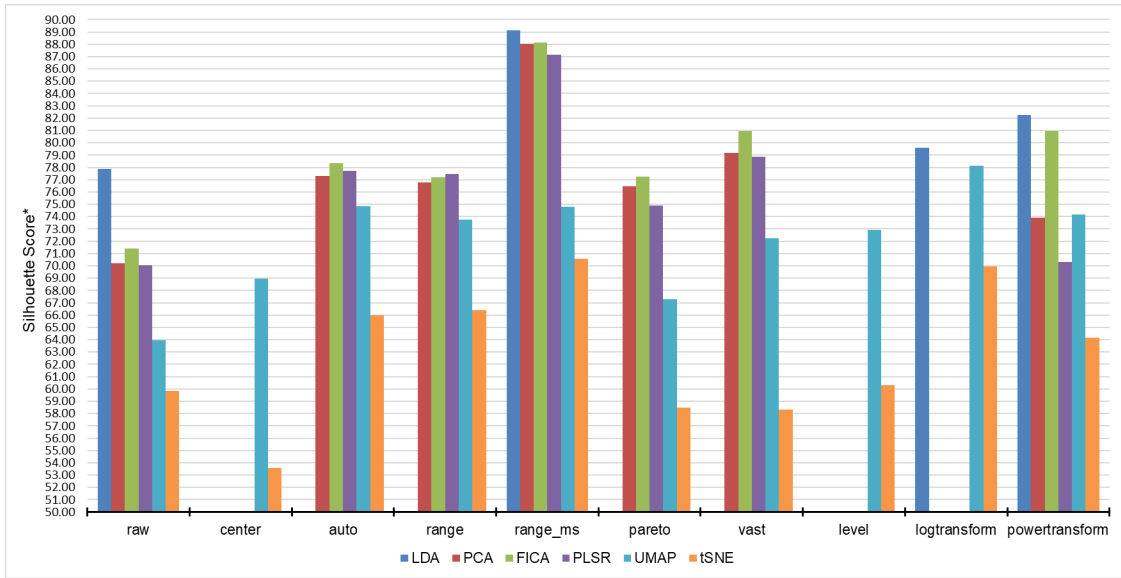


Figure 7: Average Silhouette Score* of pretreatment functions on dimensionality reduction methods.

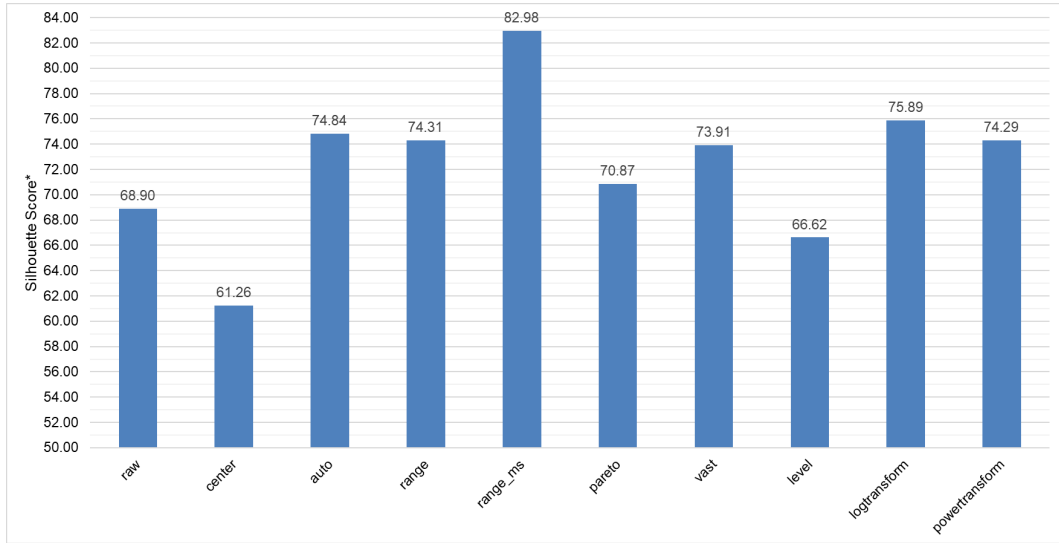


Figure 8: Global performance of pretreatment functions.

4.2. Impact of statistical measures

This section analyzes the relationship between appliance statistical measures and feature sets as well as their joint influence on dimensional reduction methods. Figure 9 denotes that the reference measure "global" obtains good results, but not the best values, which means that this statistical measure improves the overall results. However, the performance of "self" is poor. By calculating the mean values to the previous data, Figure 10 is obtained, which shows how "CUS" feature set obtains higher results than "PH" and "CH", which is also the case when "self" and "global" statistical measures are utilized.

Figure 11 compares the average performance of dimensional reduction methods after applying statistical measures substitution. In general terms, it is observed that the DR methods often show a clear decline of results for the reference statistical measures "self" and "global", although in the case of UMAP, the results obtained by using "global" are better than those obtained by other DR methods. Overall, the best performance is obtained by LDA, while PCA, PLSR and FICA also obtain acceptable results. On the contrary non-linear DR methods obtain a poor performance, especially tSNE.

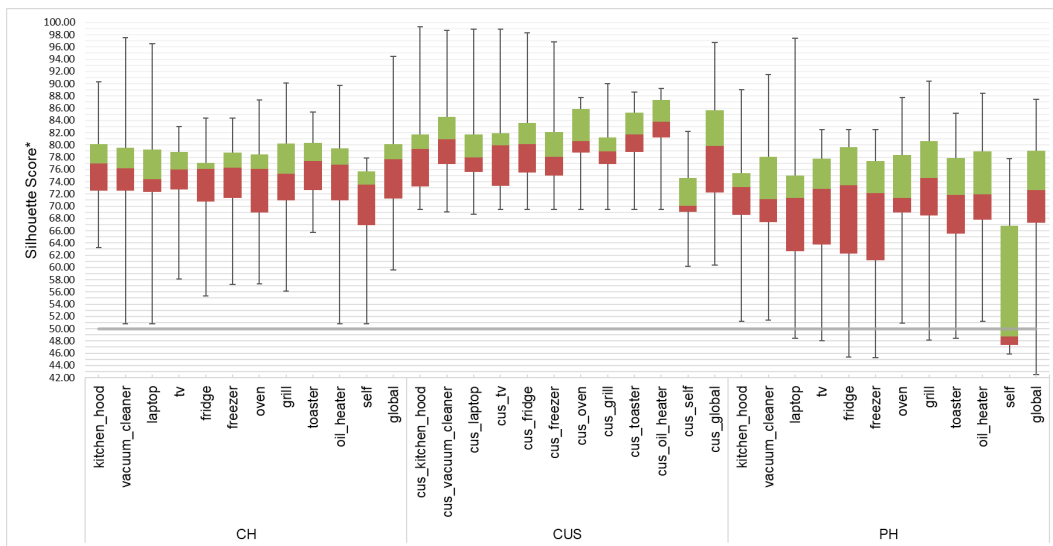


Figure 9: Effect of statistical measures in each type of feature set.

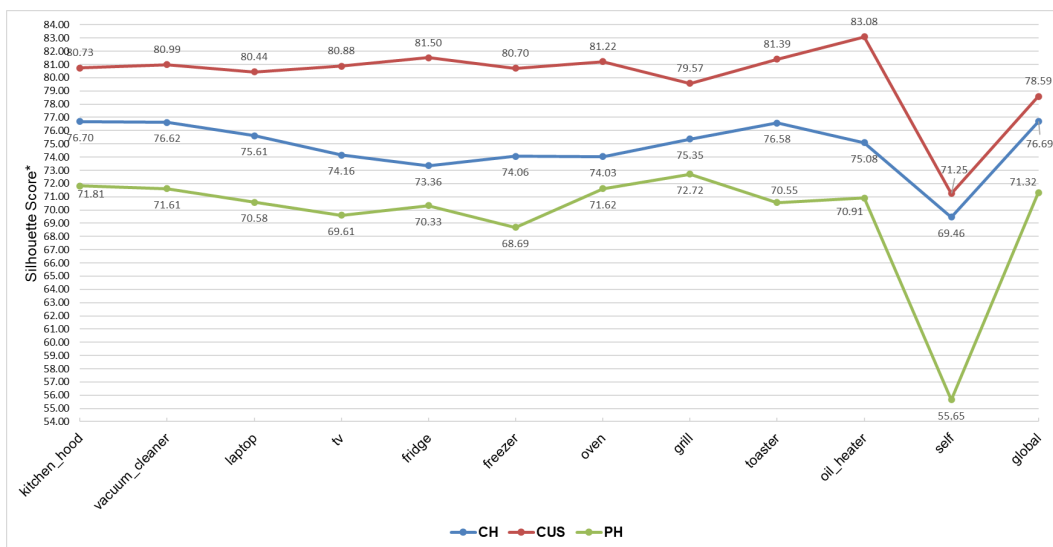


Figure 10: Impact of average appliance statistical measures substitution on feature set.

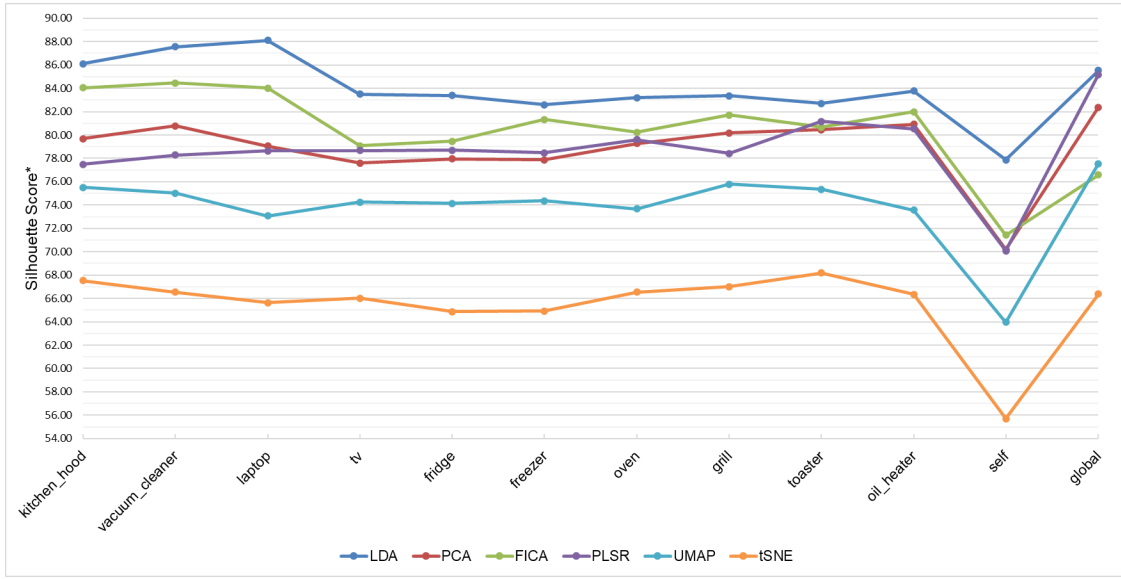


Figure 11: Impact of the DR method in different home appliances considering average values.

4.3. Impact of the type of feature set on 2D reduction methods

Figure 12 represents the result distribution of feature sets and dimensionality reduction methods. The LDA method obtains good results for all feature sets with very prominent values for the custom combination "CUS". On the opposite side is tSNE, which obtains values crossing the 50 mark, signifying overlapping clusters.

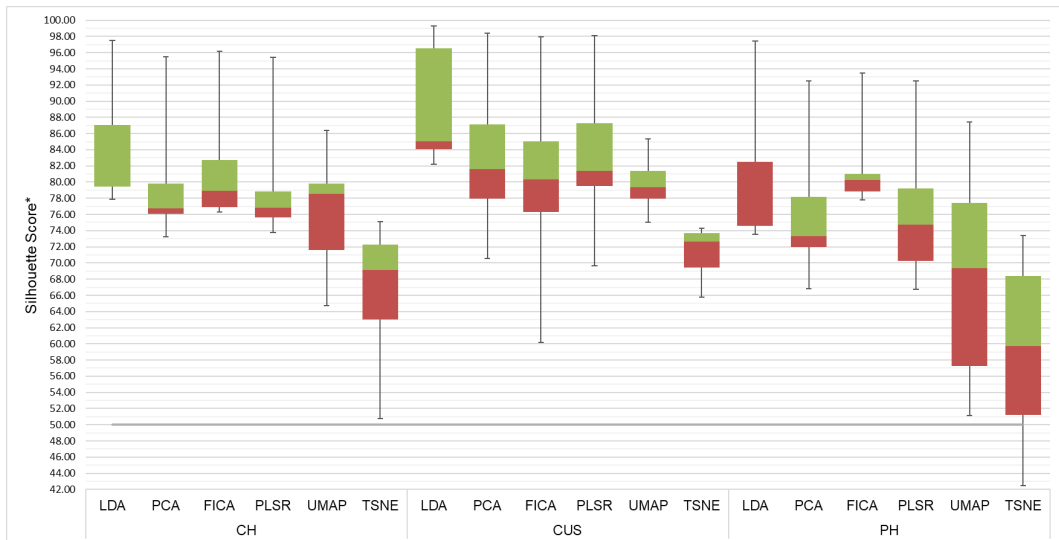


Figure 12: Impact of harmonic feature set over dimensional reduction methods.

A similar pattern is observed in Figure 13, which displays the average feature set performance. In general, the performance of the three feature sets from best to worst is "CUS", "CH" and "PH". It is worth commenting that the application of FICA provides very similar results using the three feature sets.

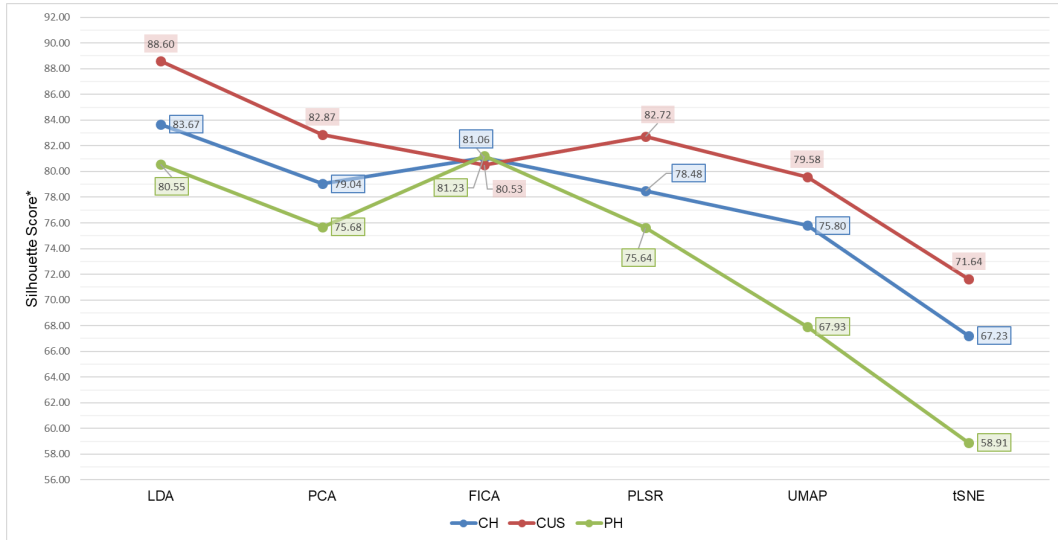


Figure 13: Average impact of harmonic feature sets over the reduction methods. Three lines have a similar shape, which resembles the average performance that offers each harmonic feature set.

4.4. Runtime analysis of dimensionality reduction methods

An important task when comparing different alternatives is to analyze the runtime required to compute them. Table 5 presents the statistical values of datasets that improved raw results. It should be noted that the times shown in Table 5 correspond to the times used by the different dimensionality reduction methods considering the data recorded by oZm during 10-minute recordings. As can be seen, the times required by the non-linear procedures (UMAP and tSNE) are significantly higher than the linear methods (LDA, PCA, FICA, PLSR).

		MAX	MEAN	MIN	StDev
CH	LDA	0.25	0.24	0.23	0.01
	PCA	0.08	0.06	0.06	0.00
	FICA	0.59	0.30	0.22	0.02
	PLSR	0.13	0.06	0.05	0.00
	UMAP	18.92	16.34	14.79	0.07
	tSNE	59.57	52.76	47.56	0.29
CUS	LDA	0.60	0.55	0.53	0.00
	PCA	0.31	0.23	0.20	0.00
	FICA	1.37	1.05	0.91	0.01
	PLSR	0.17	0.14	0.13	0.00
	UMAP	18.00	15.95	15.34	0.09
	tSNE	60.61	53.98	47.80	0.53
PH	LDA	0.25	0.23	0.18	0.00
	PCA	0.11	0.06	0.05	0.00
	FICA	0.34	0.27	0.22	0.01
	PLSR	0.11	0.05	0.05	0.00
	UMAP	29.28	17.09	14.92	0.36
	tSNE	64.27	51.97	46.03	0.40

Table 5: DR Method execution time (in seconds) statistics calculated over dataset scores that improved raw results.

4.5. Clustering plots

In addition to the previous results, they are presented some interesting figures of the results obtained for different home appliances, which associated colors are shown in Table 6. Each figure contains six plots, such that the three images at the top show the three best performing combinations for the given DR method and the feature sets, while the three images at the bottom of each figure contain the plots without any pretreatment function applied, that is, the raw data. Feature set type, pretreatment function, appliance tag and Silhouette Score* are included in these figures.











TAG:	kitchen_	vacuum_	laptop	tv	fridge	freezer	oven	grill	toaster	oil_
	hood	cleaner								heater
COLOR:										

Table 6: Appliance color legend used in the plots.

4.5.1. LDA

As described in Figure 5, the best performing pretreatment combination marked in orange color, is achieved with LDA DR method. Figures 14(d), 14(e), 14(f) show the dimensionally reduced raw data and present overlapping of the clusters which is partially or completely solved with pretreatment. In particular, Figure 14(e) shows a good score of 82.20 for untreated data using "CUS" feature set, but it reaches a very high score (99.27) after applying the pretreatment function `scale_range_ms`, as shown in Figure 14(b),

which corresponds to the best score achieved. Figure 14(a) has a high score of 97.50, but presents cluster overlapping between toaster and grill, which could be explained by the great distance of the laptop cluster which affects the Silhouette Score calculation (Equation 2). Figure 14(c) shows a very similar score (97.40) but without the overlapping problems observed in Figure 14(a).

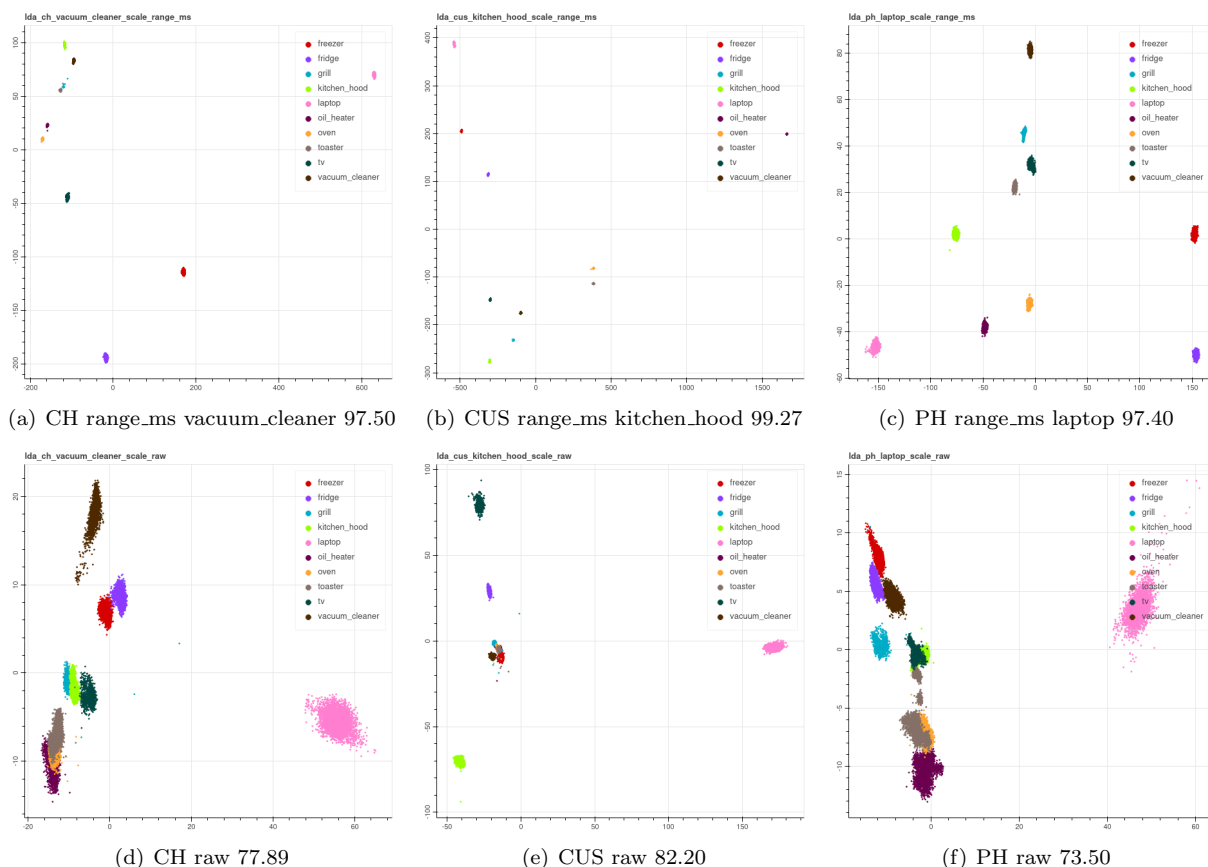


Figure 14: DR Method: LDA. Silhouette score values mapped from $[-1,1]$ to $[0,100]$. Figure labels contain Feature set, applied function, statistical measures applied, score. Figures a,b,c are the pretreated version of Figures d,e,f.

4.5.2. PCA

The comparison of Figures 15(d), 15(e), 15(f) denotes that raw data processed using "CH" feature set offers a slightly better score (72.23) than the custom harmonic combination (70.52) and power harmonics scores (66.82). However, after applying pretreatment methods, the scores are highly improved. In fact, Figure 15(b) shows that "CUS" scores 98.41, while "CH" (Figure 15(a)) and "PH" (Figure 15(c)) score 95.53 and 92.47, respectively. Interestingly, clusters tend to be grouped by the appliance type in the case of using "CH" dataset (see Figure 15(a)), since resistive appliances such as electric grill, oven, oil heater and toaster are close to each other, while appliances with a motor such as kitchen hood and vacuum cleaner are

close each other as it can be observed in the top left corner of the figure. Similarly, freezer and fridge are very close.

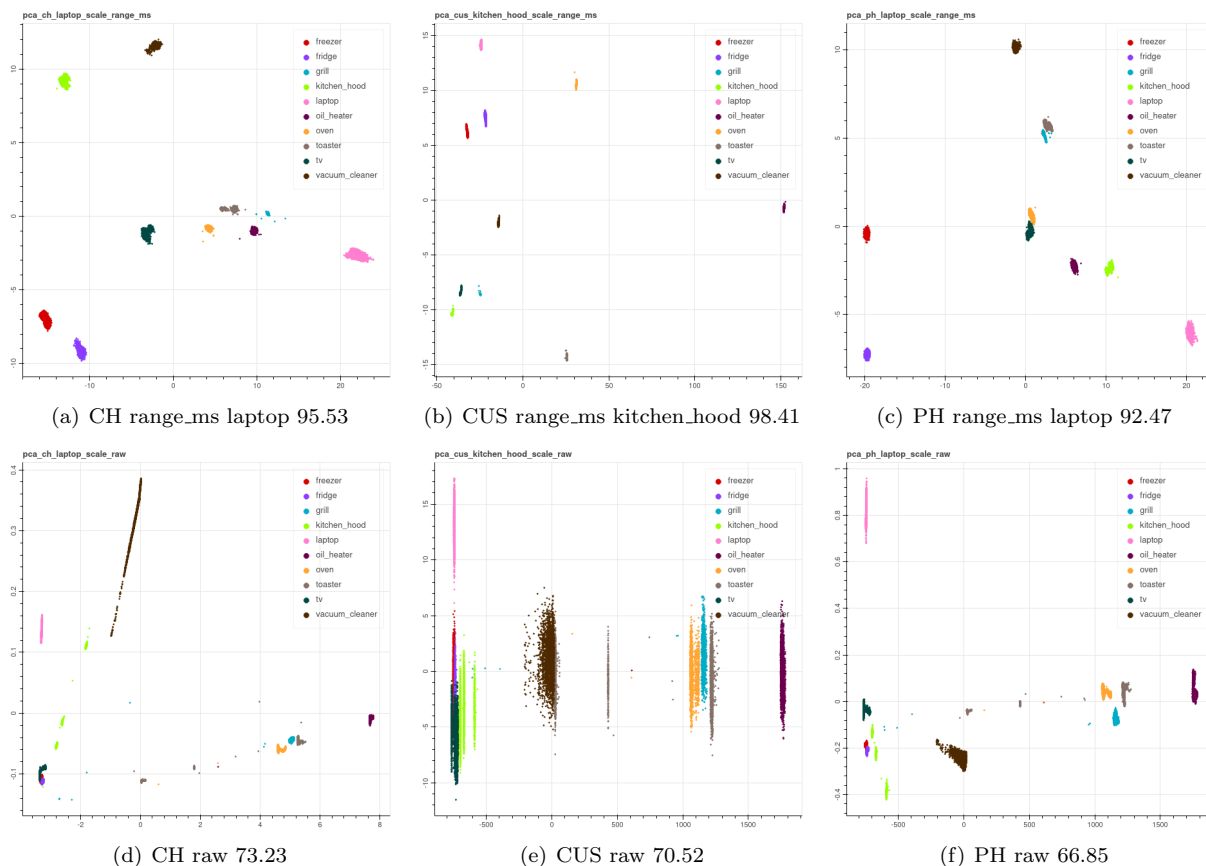


Figure 15: DR Method: PCA. Silhouette score values mapped from $[-1,1]$ to $[0,100]$. Figure labels contain Feature set, applied function, statistical measures applied, score. Figures a,b,c are the pretreated version of Figures d,e,f.

4.5.3. FICA

In the case of FICA, it is observed that current harmonic data without pretreatment (Figure 16(d)) scores 76.32, while the application of pretreatment with scale_range.ms function using vacuum_cleaner statistical measure improved the score to 96.13, although some cluster overlapping is observed in Figure 16(a). Power harmonics (Figure 16(f)) provide more compact clusters than Current Harmonics using raw data. The improvement can be noticed in the combined scattered data points of the toaster and other appliances displayed in Figure 16(c). Laptop statistical measures together with range.ms function seem to have a positive effect on raw data, which scores 93.48 in the case of FICA DR method. The best pretreatment performance is observed using "CUS" dataset, such that using range.ms function allows the score to increase from 60.21 when raw data is considered (Figure 16(e)) to 97.98 with compact and well separated clusters as

Figure 16(b) shows.

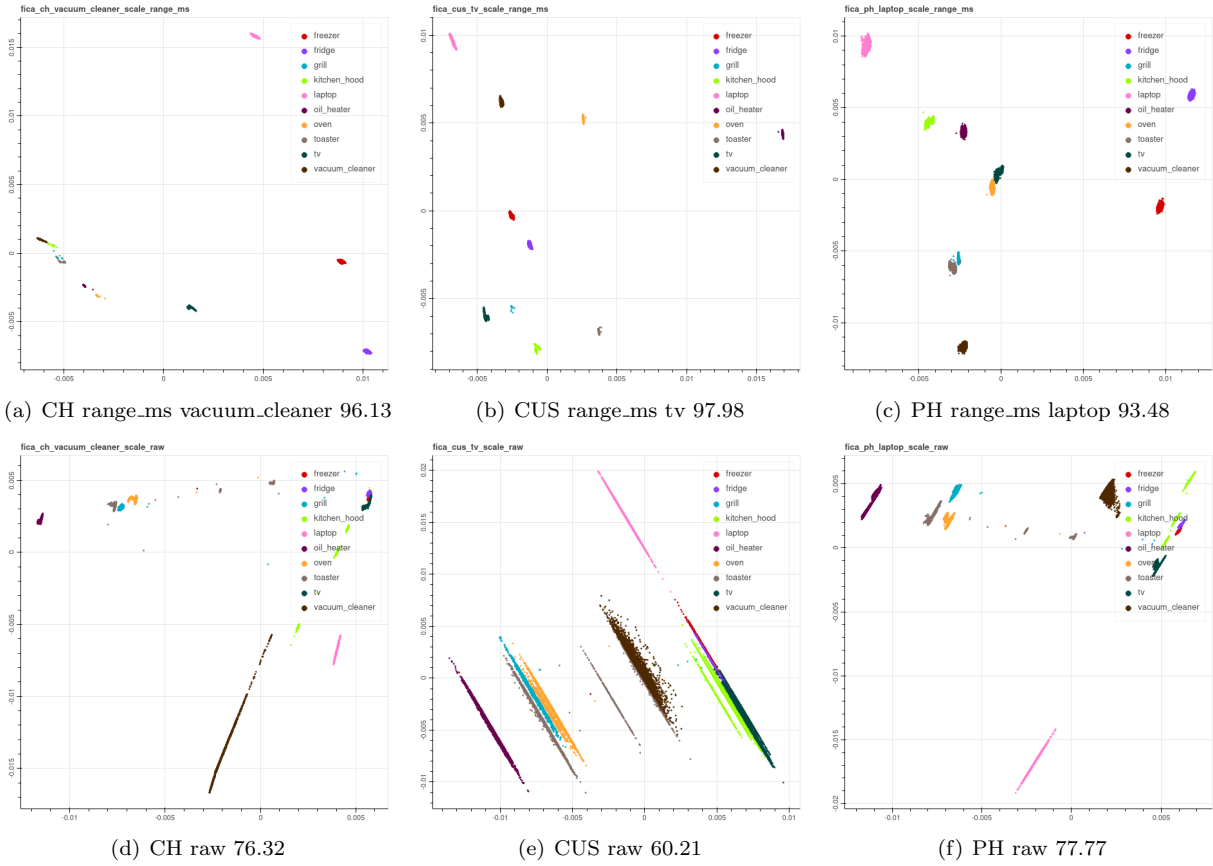


Figure 16: DR Method: FICA. Silhouette score values mapped from $[-1,1]$ to $[0,100]$. Figure labels contain Feature set, applied function, statistical measures applied, score. Figures a,b,c are the pretreated version of Figures d,e,f

4.5.4. PLSR

PLSR figures of raw data present similar cluster shapes to those of FICA DR method. Current harmonic raw data scores 73.80 ((Figure 17(d)), while the score is 95.40 when vacuum cleaner statistical measures and scale_range_ms are applied (see Figure 17(a)). Custom harmonics (Figure 17(e)) displays clusters with line shapes and a score of 69.67, after pretreatment compact and separated clusters are obtained, with a score of 98.08. Figure 17(b) shows the result of applying scale_range_ms function with kitchen hood statistical measures. Improvement of power harmonics raw (Figure 17(f)) is made evident by the increase in score from 66.75 to 92.51. Figure 17(c) shows that scale range_ms function with laptop statistical measures compacts and separates data points.

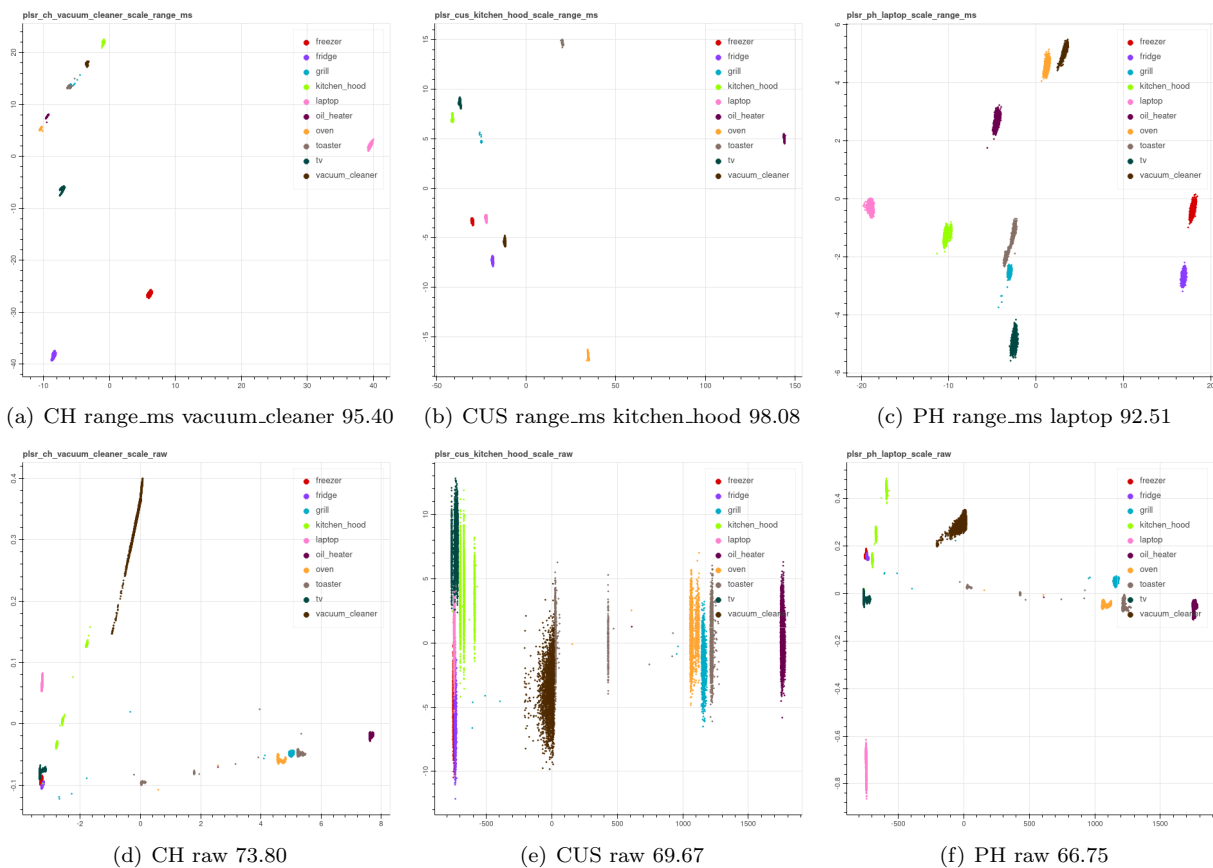


Figure 17: DR Method: PLSR. Silhouette score values mapped from $[-1,1]$ to $[0,100]$. Figure labels contain Feature set, applied function, statistical measures applied, score. Figures a,b,c are the pretreated version of Figures d,e,f

4.5.5. UMAP

In the case of UMAP, the scores for raw and pretreated data are not as high as for the previous linear methods. Current harmonics raw score is 68.24 (Figure 18(d)) and provides some overlapping and scattered clusters, while the application of scale_range and a global statistical measure allows a score of 86.35 to be obtained. Custom combination raw score is 78.60 (Figure 18(e)), and pretreatment improves the score (85.37) by applying Pareto function and global measures. The comparison of Figure 18(c) with Figure 18(f) shows that power harmonics raw data obtains the worst score (53.02), but it is significantly improved by using range function with global measure, providing a resulting score of 87.45.

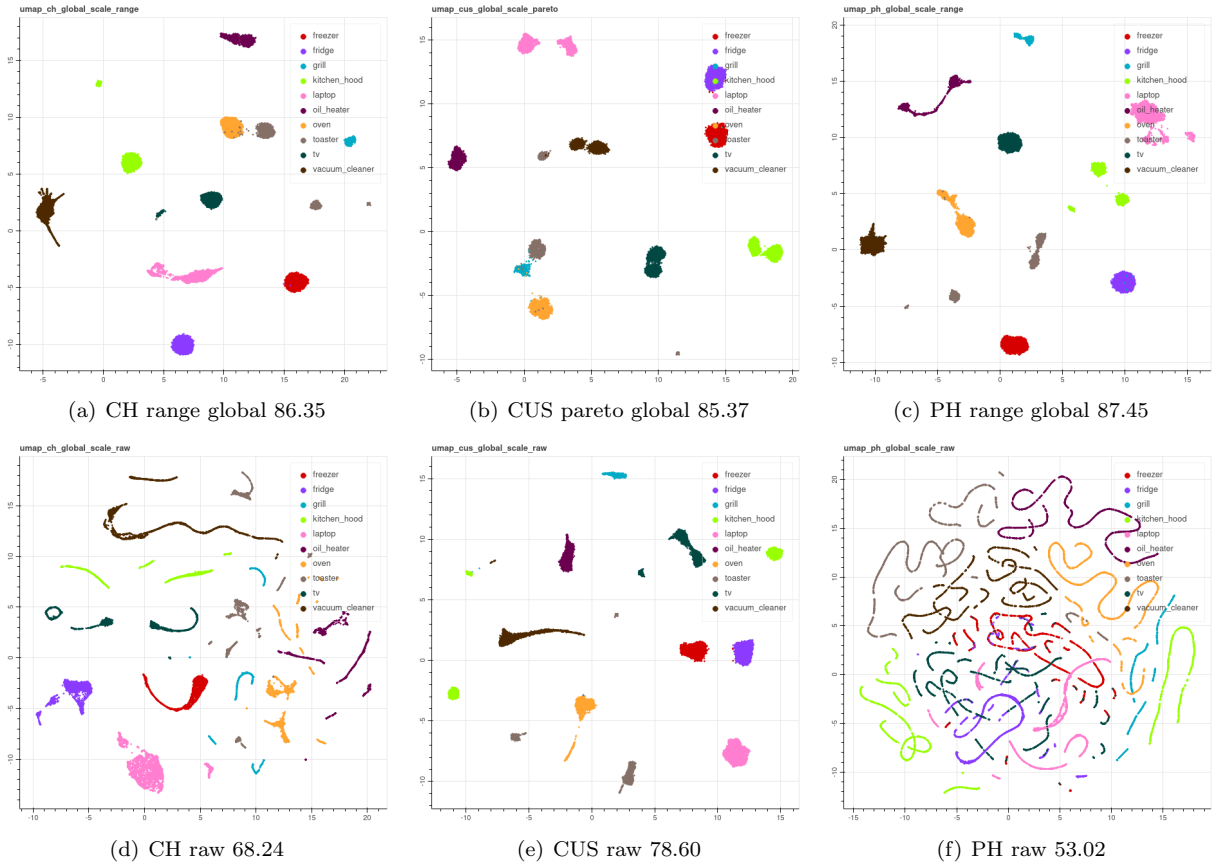


Figure 18: DR Method: UMAP. Silhouette score values mapped from $[-1,1]$ to $[0,100]$. Figure labels contain Feature set, applied function, statistical measures applied, score. Figures a,b,c are the pretreated version of Figures d,e,f

4.5.6. tSNE

TSNE and UMAP are similar methods and the results are very alike for the good and the bad results. For example, Power harmonics raw data (see Figure 19(f)) provides a poor score of 45.30 (below 0 on the original Silhouette range), but the application of scale_range function with grill statistical measures improved the DR process and helped in forming clusters with reasonable shapes and a score of 73.41, as Figure 19(c) shows. Current harmonic raw data (Figure 19(d)) scores 50.79 and contains scattered cluster shapes all over the plot. In this case, the use of scale_range function with laptop measures improves the score to 75.12, but clusters are not fully connected and compact (Figure 19(a)). Finally, "CUS" dataset using raw data (Figure 19(e)) scores 65.58 and even though the score increases to 74.31 using pretreated functions, it presents more point scattering and overlapping between clusters (see Figure 19(b)).

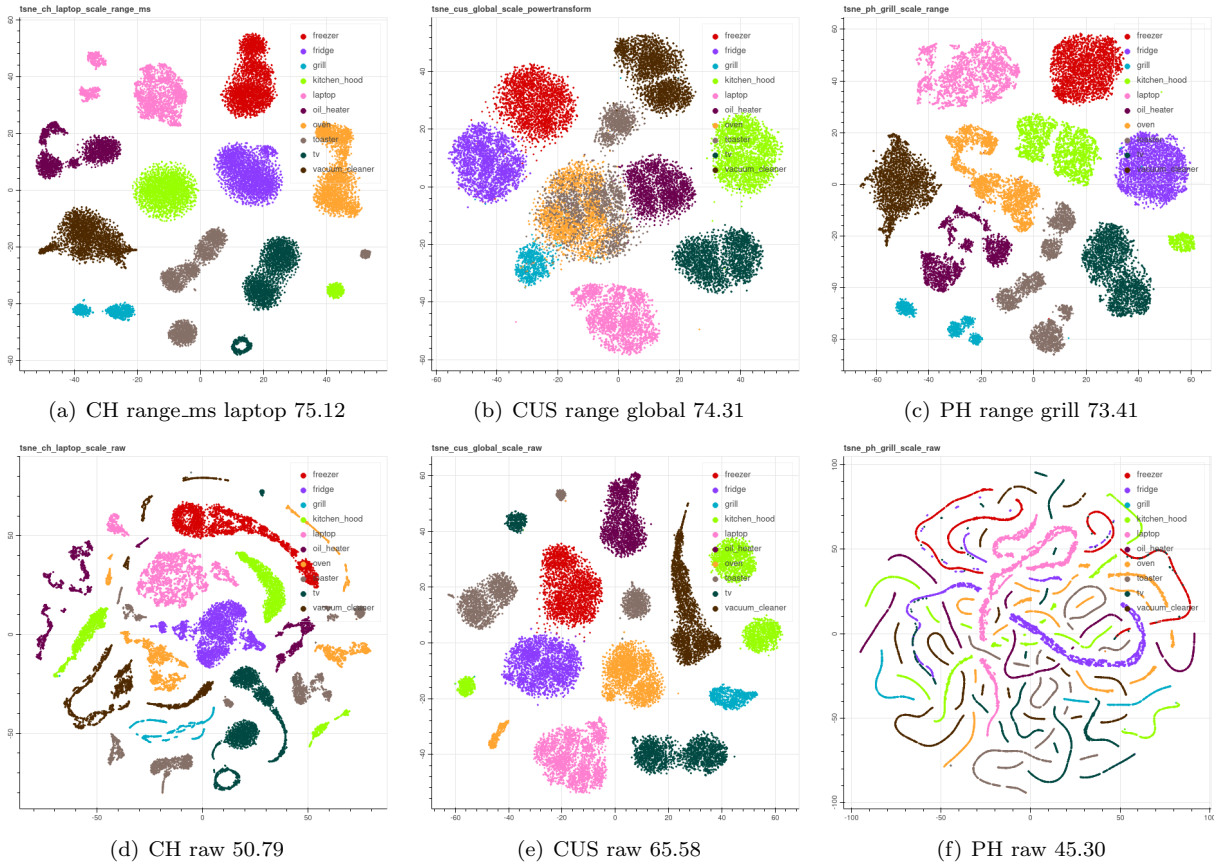


Figure 19: DR Method: tSNE. Silhouette score values mapped from $[-1,1]$ to $[0,100]$. Figure labels contain Feature set, applied function, statistical measures applied, score. Figures a,b,c are the pretreated version of Figures d,e,f

4.6. Explained variance

Explained variance [70] is a metric for gauging the disparity between the output of a given method and the actual data. Higher percentages of explained variance indicates a greater strength of association. Some previous studies have used the explained variance to determine the performance of dimensionality reduction methods. For example, in [71] the cumulative and normalized eigenvalues for PCA are calculated in order to represent the percentage of explained total variance.

Figure 20 presents the sum of the explained variance considering the two first components for PCA and LDA. The same heatmap criteria considered in Figure 5 is also used in Figure 20. As it can be observed, these results show that the use of two components allows us to obtain a high value of explained variance.

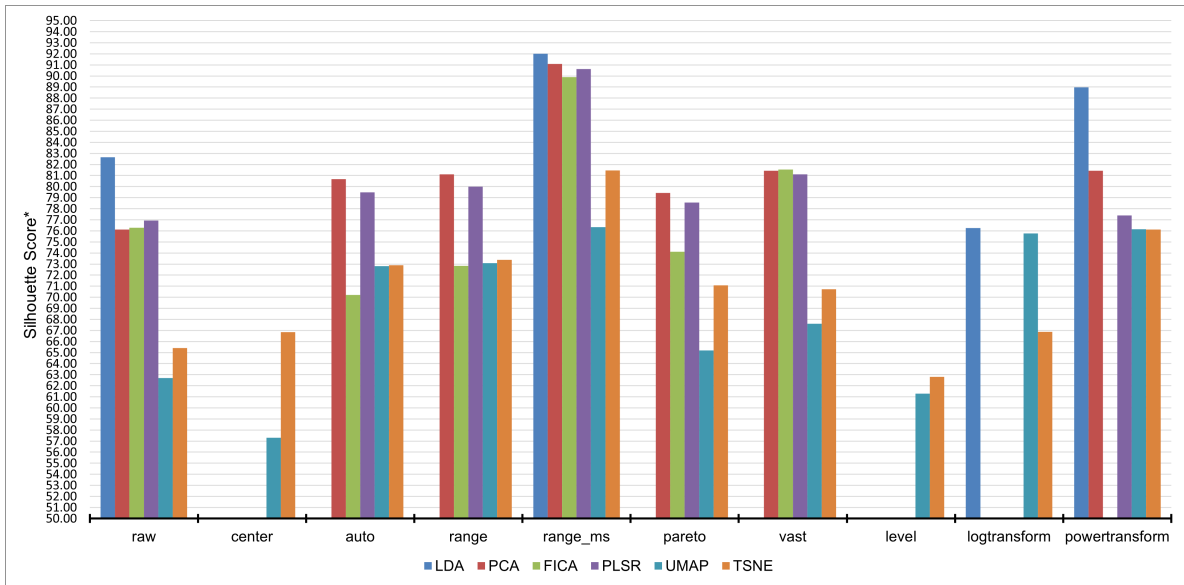


Figure 21: Average pretreatment function impact on 2D reduction methods (COOLL dataset).

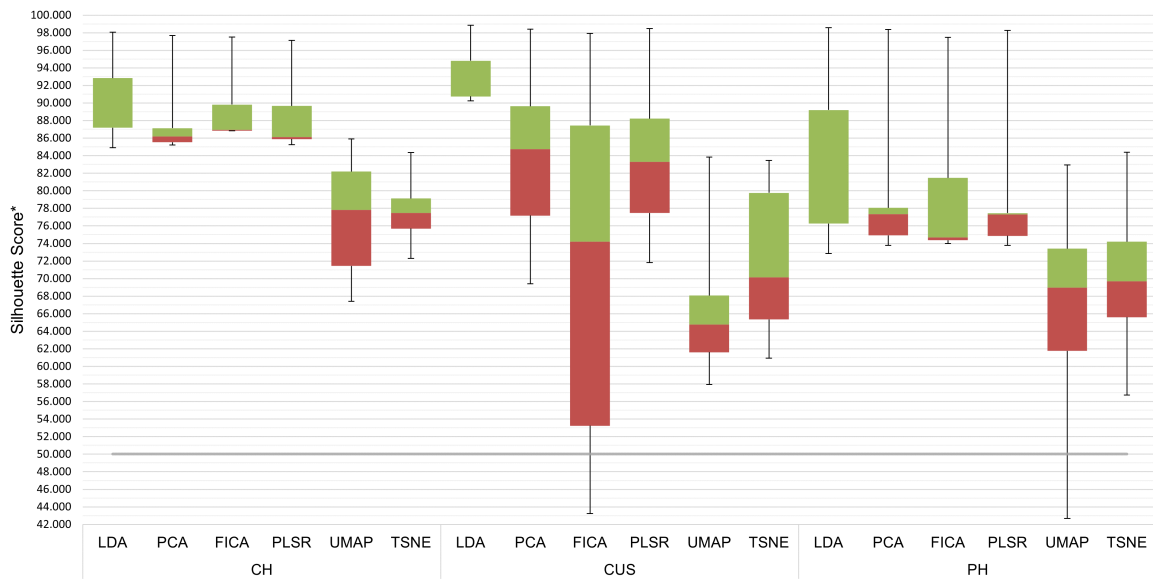


Figure 22: Impact of harmonic feature set over dimensional reduction methods (COOLL dataset).

5. Discussion and conclusions

This paper analyses the performance of data pretreatment techniques and dimensionality reduction methods on energy consumption and power quality data measured by an advanced metering infrastructure in

a real environment composed of different household appliances. Specifically, 8 data pretreatment techniques and 6 dimensionality reduction methods have been implemented and compared on energy consumption and power quality data measured by an advanced metering infrastructure considering the operation of 10 different household appliances. The data has been obtained considering long-term (10 minutes) data recording strategy of set of household appliances. In addition, the procedure presented here has been successfully applied with public databases containing power consumption profiles of a wide variety of devices. The results obtained clearly show that the information provided by the raw data can be improved by using pretreatment techniques and dimensionality reduction methods. In particular, the results obtained by the *range_ms* function obtains the best results among the eight pretreatment methods considered here. As for the dimensionality reduction methods, it is observed that linear approaches outperform non-linear ones, with Linear Discriminant Analysis (LDA) obtaining the best overall result.

The methodology presented in this article, as well as the results obtained with real data from different household appliances, have important implications for the purposes of NILM. In particular, this research contributes to the analysis and development of advanced dimensionality reduction techniques and to the detection of possible linear dependencies between variables. A contribution of great interest is the analysis of current and active power harmonics to carry out the process, such that it has been shown that the customised combination of active power and current harmonics can be successfully employed to obtain improvements in the quality of the solutions found. The results obtained can be quite useful for any researcher interested in disaggregating information contained in voltage and current signals of household appliances or any other load consuming electricity. Moreover, when evaluated on well-known datasets, the suggested strategy performs well in terms of standard measures and time complexity and may be utilized with any NILM classification algorithm. An important limitation of this study lies in the fact that some appliances include several internal loads (e.g. washing machines include water drainage pumps, heating elements, motors, etc.), reason why in the future it is planned to study these type of appliances.

Acknowledgment

This research has been supported by the Spanish Ministry of Science, Innovation and Universities under the programme *Proyectos de I+D de Generación de Conocimiento* of the research, development and innovation system with grant number PGC2018-098813-B-C33.

References

- [1] Y. Lu, N. Nakicenovic, M. Visbeck, A.-S. Stevance, Policy: Five priorities for the UN sustainable development goals, *Nature News* 520 (7548) (2015) 432.
- [2] A. Zakari, I. Khan, D. Tan, R. Alvarado, V. Dagar, Energy efficiency and sustainable development goals (sdgs), *Energy* 239 (2022) 122365.
- [3] S. Aheleroff, X. Xu, Y. Lu, M. Aristizabal, J. P. Velásquez, B. Joa, Y. Valencia, Iot-enabled smart appliances under industry 4.0: A case study, *Advanced Engineering Informatics* 43 (2020) 101043.
- [4] H. Akhavan-Hejazi, H. Mohsenian-Rad, Power systems big data analytics: An assessment of paradigm shift barriers and prospects, *Energy Reports* 4 (2018) 91–100.
- [5] K. Moharm, State of the art in big data applications in microgrid: A review, *Advanced Engineering Informatics* 42 (2019) 100945.
- [6] S. Yilmaz, S. K. Firth, D. Allinson, Occupant behaviour modelling in domestic buildings: the case of household electrical appliances, *Journal of Building Performance Simulation* 10 (5-6) (2017) 582–600.
- [7] A. J. Sonta, P. E. Simmons, R. K. Jain, Understanding building occupant activities at scale: An integrated knowledge-based and data-driven approach, *Advanced Engineering Informatics* 37 (2018) 1–13.
- [8] A. Ridi, C. Gisler, J. Hennebert, A survey on intrusive load monitoring for appliance recognition, in: 2014 22nd international conference on pattern recognition, IEEE, 2014, pp. 3702–3707.
- [9] G. W. Hart, Nonintrusive appliance load monitoring, *Proceedings of the IEEE* 80 (12) (1992) 1870–1891.
- [10] S. Giri, M. Bergés, An error correction framework for sequences resulting from known state-transition models in non-intrusive load monitoring, *Advanced Engineering Informatics* 32 (2017) 152–162.
- [11] J. Alcalá, J. Ureña, Á. Hernández, D. Gualda, Event-based energy disaggregation algorithm for activity monitoring from a single-point sensor, *IEEE Transactions on Instrumentation and Measurement* 66 (10) (2017) 2615–2626.
- [12] G. Bedi, G. K. Venayagamoorthy, R. Singh, R. R. Brooks, K.-C. Wang, Review of Internet of Things (IoT) in electric power and energy systems, *IEEE Internet of Things Journal* 5 (2) (2018) 847–870.
- [13] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du, S. Marshall, Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging, *Neurocomputing* 185 (2016) 1–10.
- [14] E. Esser, M. Moller, S. Osher, G. Sapiro, J. Xin, A convex model for nonnegative matrix factorization and dimensionality reduction on physical space, *IEEE Transactions on Image Processing* 21 (7) (2012) 3239–3252.
- [15] F. Anowar, S. Sadaoui, B. Selim, Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne), *Computer Science Review* 40 (2021) 100378.
- [16] A. Morán, J. J. Fuertes, M. A. Prada, S. Alonso, P. Barrientos, I. Díaz, M. Domínguez, Analysis of electricity consumption profiles in public buildings with dimensionality reduction techniques, *Engineering Applications of Artificial Intelligence* 26 (8) (2013) 1872–1880.
- [17] M. Khodayar, G. Liu, J. Wang, M. E. Khodayar, Deep learning in power systems research: A review, *CSEE Journal of Power and Energy Systems* 7 (2) (2020) 209–220.
- [18] A. S. Bouhouras, P. A. Gkaidatzis, E. Panagiotou, N. Poulakis, G. C. Christoforidis, A NILM algorithm with enhanced disaggregation scheme under harmonic current vectors, *Energy and Buildings* 183 (2019) 392–407.
- [19] H. Kang, H. Kim, et al., Household appliance classification using lower odd-numbered harmonics and the bagging decision tree, *IEEE Access* 8 (2020) 55937–55952.

- [20] R. Machlev, D. Tolkachov, Y. Levron, Y. Beck, Dimension reduction for nilm classification based on principle component analysis, *Electric Power Systems Research* 187 (2020) 106459.
- [21] C. L. Athanasiadis, T. A. Papadopoulos, D. I. Doukas, Real-time non-intrusive load monitoring: A light-weight and scalable approach, *Energy and Buildings* 253 (2021) 111523.
- [22] E. Viciano, A. Alcayde, F. G. Montoya, R. Baños, F. M. Arrabal-Campos, F. Manzano-Agugliaro, An open hardware design for internet of things power quality and energy saving solutions, *Sensors* 19 (3) (2019) 627.
- [23] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, Centering, scaling, and transformations: Improving the biological information content of metabolomics data, *BMC Genomics* 7 (1) (2006) 1–15.
- [24] Q. Huang, H. Cheng, K. Fang, Y. Huang, T. Liu, S. Yang, Research on engineering application of the nonintrusive load monitoring technology, in: *2021 IEEE 4th International Conference on Electronics Technology (ICET)*, IEEE, 2021, pp. 504–508.
- [25] A. Borrmann, P. Geyer, C. Koch, Advanced computing for the built environment, *Advanced Engineering Informatics* 27 (4) (2013) 411–412.
- [26] Y. Himeur, A. Alsalemi, F. Bensaali, A. Amira, An intelligent nonintrusive load monitoring scheme based on 2d phase encoding of power signals, *International Journal of Intelligent Systems* 36 (1) (2021) 72–93.
- [27] W. A. Souza, A. M. Alonso, T. B. Bosco, F. D. Garcia, F. A. Gonçalves, F. P. Marafão, Selection of features from power theories to compose nilm datasets, *Advanced Engineering Informatics* 52 (2022) 101556.
- [28] J. Revuelta Herrero, Á. Lozano Murciego, A. López Barriuso, D. Hernández de la Iglesia, G. Villarrubia González, J. M. Corchado Rodríguez, R. Carreira, Non intrusive load monitoring (nilm): A state of the art, in: *International Conference on Practical Applications of Agents and Multi-Agent Systems*, Springer, 2017, pp. 125–138.
- [29] B. Kalluri, A. Kamilaris, S. Kondepudi, H. W. Kua, K. W. Tham, Applicability of using time series subsequences to study office plug load appliances, *Energy and Buildings* 127 (2016) 399–410.
- [30] N. Jin, F. Yang, Y. Mo, Y. Zeng, X. Zhou, K. Yan, X. Ma, Highly accurate energy consumption forecasting model based on parallel lstm neural networks, *Advanced Engineering Informatics* 51 (2022) 101442.
- [31] B. Raducanu, F. Dornaika, A supervised non-linear dimensionality reduction approach for manifold learning, *Pattern Recognition* 45 (6) (2012) 2432–2444.
- [32] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, J. Saeed, A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction, *Journal of Applied Science and Technology Trends* 1 (2) (2020) 56–70.
- [33] G.-F. Angelis, C. Timplalexis, S. Krinidis, D. Ioannidis, D. Tzovaras, Nilm applications: Literature review of learning approaches, recent developments and challenges, *Energy and Buildings* (2022) 111951.
- [34] A. Langevin, M.-A. Carbonneau, M. Cheriet, G. Gagnon, Energy disaggregation using variational autoencoders, *Energy and Buildings* 254 (2022) 111623.
- [35] W. Kong, Z. Y. Dong, B. Wang, J. Zhao, J. Huang, A practical solution for non-intrusive type ii load monitoring based on deep learning and post-processing, *IEEE Transactions on Smart Grid* 11 (1) (2020) 148–160.
- [36] Z. Fang, D. Zhao, C. Chen, Y. Li, Y. Tian, Nonintrusive appliance identification with appliance-specific networks, *IEEE Transactions on Industry Applications* 56 (4) (2020) 3443–3452.
- [37] Y. Himeur, A. Alsalemi, F. Bensaali, A. Amira, Smart non-intrusive appliance identification using a novel local power histogramming descriptor with an improved k-nearest neighbors classifier, *Sustainable Cities and Society* 67 (2021) 102764.
- [38] S.-H. Yi, J. Wang, J.-J. Liu, Simultaneous load identification method based on hybrid features and genetic algorithm for

- nonintrusive load monitoring, *Mathematical Problems in Engineering* 2022.
- [39] F. Jazizadeh, B. Becerik-Gerber, M. Berges, L. Soibelman, An unsupervised hierarchical clustering based heuristic algorithm for facilitated training of electricity consumption disaggregation systems, *Advanced Engineering Informatics* 28 (4) (2014) 311–326.
- [40] S. Giri, M. Bergés, A. Rowe, Towards automated appliance recognition using an emf sensor in nilm platforms, *Advanced Engineering Informatics* 27 (4) (2013) 477–485.
- [41] R. Reddy, V. Garg, V. Pudi, A feature fusion technique for improved non-intrusive load monitoring, *Energy Informatics* 3 (1) (2020) 1–15.
- [42] R. Ramadan, Q. Huang, O. Bamisile, A. S. Zalhaf, Intelligent home energy management using internet of things platform based on nilm technique, *Sustainable Energy, Grids and Networks* (2022) 100785.
- [43] Y. Jimenez, C. Duarte, J. Petit, J. Meyer, P. Schegner, G. Carrillo, Steady state signatures in the time domain for nonintrusive appliance identification, *Ingeniería e Investigación* 35 (2015) 58–64.
- [44] C. Laughman, K. Lee, R. Cox, S. Shaw, S. Leeb, L. Norford, P. Armstrong, Power signature analysis, *IEEE power and energy magazine* 1 (2) (2003) 56–63.
- [45] J. Kolter, M. Johnson, The reference energy disaggregation data set (2011).
URL <http://redd.csail.mit.edu/>
- [46] J. Kelly, W. Knottenbelt, The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes, *Scientific data* 2 (150007) (2015) 1–14.
URL <https://jack-kelly.com/data/>
- [47] N. Batra, M. Gulati, A. Singh, M. B. Srivastava, It's different: Insights into home energy consumption in india, in: *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, 2013, pp. 1–8.
- [48] D. P. B. Renaux, F. Pottker, H. C. Ancelmo, A. E. Lazzaretti, C. R. E. Lima, R. R. Linhares, E. Oroski, L. d. S. Nolasco, L. T. Lima, B. M. Mulinari, et al., A dataset for non-intrusive load monitoring: Design and implementation, *Energies* 13 (20) (2020) 5371.
- [49] H. K. Iqbal, F. H. Malik, A. Muhammad, M. A. Qureshi, M. N. Abbasi, A. R. Chishti, A critical review of state-of-the-art non-intrusive load monitoring datasets, *Electric Power Systems Research* 192 (2021) 106921.
- [50] F. Jazizadeh, M. Afzalan, B. Becerik-Gerber, L. Soibelman, Embed: A dataset for energy monitoring through building electricity disaggregation, in: *Proceedings of the Ninth International Conference on Future Energy Systems*, 2018, pp. 230–235.
- [51] D. Djenouri, R. Laidi, Y. Djenouri, I. Balasingham, Machine learning for smart building applications: Review and taxonomy, *ACM Computing Surveys (CSUR)* 52 (2) (2019) 1–36.
- [52] K. J. Millman, M. Aivazis, Python for scientists and engineers, *Computing in Science & Engineering* 13 (2) (2011) 9–12.
- [53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [54] H. Kim, H. K. Kim, S. Cho, Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling, *Expert Systems with Applications* 150 (2020) 113288.
- [55] J. Liang, S. K. Ng, G. Kendall, J. W. Cheng, Load signature study—part i: Basic concept, structure, and methodology, *IEEE transactions on power Delivery* 25 (2) (2009) 551–560.

- [56] B. Worley, R. Powers, Multivariate analysis in metabolomics, *Current Metabolomics* 1 (1) (2013) 92–107.
- [57] U. Roessner, J. Bowne, What is metabolomics all about?, *Biotechniques* 46 (5) (2009) 363–365.
- [58] H. C. Keun, T. M. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, E. Holmes, J. C. Lindon, J. K. Nicholson, Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling, *Analytica Chimica Acta* 490 (1-2) (2003) 265–276.
- [59] F. Changyong, W. Hongyue, L. Naiji, C. Tian, H. Hua, L. Ying, et al., Log-transformation and its implications for data analysis, *Shanghai Archives of Psychiatry* 26 (2) (2014) 105.
- [60] M. S. Klein, Affine transformation of negative values for NMR metabolomics using the `mrbin` R Package, *Journal of Proteome Research* 20 (2) (2021) 1397–1404.
- [61] I. T. Jolliffe, J. Cadima, Principal component analysis: A review and recent developments, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2065) (2016) 20150202.
- [62] W. A. Souza, F. P. Marafão, E. V. Liberado, M. G. Simões, L. C. Da Silva, A `nilm` dataset for cognitive meters based on conservative power theory and pattern recognition techniques, *Journal of Control, Automation and Electrical Systems* 29 (6) (2018) 742–755.
- [63] C. H. Park, H. Park, A comparison of generalized linear discriminant analysis algorithms, *Pattern Recognition* 41 (3) (2008) 1083–1097.
- [64] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, *Neural Networks* 13 (4-5) (2000) 411–430.
- [65] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS regression), *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (1) (2010) 97–106.
- [66] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research* 9 (11) (2008) 2579–2605.
- [67] G. Hinton, S. T. Roweis, Stochastic neighbor embedding, in: *Advances in Neural Information Processing Systems*, Vol. 15, Citeseer, 2002, pp. 833–840.
- [68] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426.
- [69] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. Pérez, I. Perona, An extensive comparative study of cluster validity indices, *Pattern Recognition* 46 (2013) 243–256.
- [70] R. M. Warner, *Applied statistics: From bivariate through multivariate techniques*, Sage Publications, 2012.
- [71] D. Malmgren-Hansen, V. Laparra, A. A. Nielsen, G. Camps-Valls, Spatial noise-aware temperature retrieval from infrared sounder data, in: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2017, pp. 17–20.
- [72] T. Picon, M. N. Meziane, P. Ravier, G. Lamarque, C. Novello, J.-C. L. Bunetel, Y. Raingeaud, `Cooll`: Controlled on/off loads library, a public dataset of high-sampled electrical signals for appliance identification, arXiv preprint arXiv:1611.05803.
- URL <https://coolldataset.github.io/>