Synthetic Ground Truth Generation of an Electricity Consumption Dataset

Synthetic Ground Truth Generation of an Electricity Consumption Dataset

Lorenzo Mascali*, Simone Eiraudo*, Luca Barbierato*, Daniele Salvatore Schiera*, Roberta Giannantonio†, Edoardo Patti*, Lorenzo Bottaccioli*, and Andrea Lanzini*
*Energy Center Lab, Politecnico di Torino, Turin, Italy. Email: name.surname@polito.it
†Data Office, TIM S.p.A., Italy, Email: name.surname@telecomitalia.it

***Abstract*—The training of supervised Machine Learning (ML) and Artificial Intelligence (AI) algorithms is strongly affected by the goodness of the input data. To this end, this paper proposes an innovative synthetic ground truth generation algorithm. The methodology is based on applying a data reduction with Symbolic Aggregate Approximation (SAX). In addition, a Classification And Regression Tree (CART) is employed to identify the best granularity of the data reduction. The proposed algorithm has been applied to telecommunication (TLC) sites dataset by analyzing their electricity consumption patterns. The presented approach substantially reduced the dispersion of the dataset compared to the raw dataset, thus reducing the effort required to train the supervised algorithms.**

***Index Terms*—Approximation, Classification, Clustering, Symbols Aggregated, Synthetic Ground Truth**

NOMENCLATURE

Acronyms

AI	Artificial Intelligence
ARIMA	Auto-Regressive Integrated Moving Average
BMS	Building Management System
CART	Classification And Regression Trees
CO	Central Offices
MIA	Mean Index Adequacy
ML	Machine Learning
PCA	Principal Component Analysis
PDF	Probability Density Function
PUE	Power Usage Effectiveness
SAX	Symbolic Aggregate Approximation
SOM	Self-Organizing Map
TLC	TeLeCommunication
UF	Utilization Factor
UPS	Uninterrupted Power Supply

I. INTRODUCTION

Buildings account for a key share of the worldwide energy consumption, being responsible for more than 10000 TWh [1] of electrical energy demand per year. Specifically, commercial and industrial buildings in 2018 were responsible for 8% of total energy consumption [2]. For this reason, energy efficiency and rational resources management in this sector are significant challenges to be addressed to reduce energy footprint. In order to deal with the above-mentioned issues, academic efforts focus on two main approaches: *i)* the retrofit of building envelope and systems, and *ii)* the optimisation in the management of the existing installations. Katipamula et al. [3] show that a significant part of buildings' energy inefficiencies is caused by poor energy management rather than by inefficient building envelopes and systems. Although many buildings are equipped with a Building Management System (BMS) with distributed sensors and high sampling rates, monitored data are not always properly exploited. On the other hand, many residential, commercial, and industrial sites have not been equipped with pervasive sensors. Therefore, the research community and energy managers may rely solely on data from meters to extend the analysis to these buildings. Effective data exploitation is more and more frequently achieved by means of Machine Learning (ML) and Artificial Intelligence (AI) algorithms. These algorithms can be divided into two main categories: *i)* unsupervised and *ii)* supervised algorithms. Unsupervised algorithms allow the analysis of unclassified data and are mainly used to analyse datasets by looking for previously unknown information. Vice versa, supervised algorithms can be used whereas a dataset is already provided with some additional information, in order to mark out relationships within input and outputs. This process is also known as tagging. These algorithms can provide valuable outcomes as they are fed with extensive and reliable datasets and are adequately trained. Nevertheless, incorrect data can determine incorrect models.

The non-trivial task of defining a ground truth for training these models represents the most critical challenge of the pre-processing step in ML. Indeed, a large amount of data to be analysed discourages manual tagging of anomalies by domain experts. Therefore, the adoption of unsupervised algorithms can enhance the effective and automated extraction of reliable synthetic ground truth [4].

This paper presents a novel ML-based approach to retrieve synthetic ground truth from an electricity consumption dataset. This methodology is based on the application of an optimised Symbolic Aggregate Approximation (SAX) algorithm for data reduction and a Classification And Regression Tree (CART). The optimal setup of the employed tool is achieved by considering many configurations and comparing their Mean Index Adequacy (MIA). By applying the methodology, the algorithm allows clustering of the dataset according to the likeliness of the buildings' daily consumption profile [5]. Then, the most significant clusters are selected to form the synthetic ground truth by applying the Pareto principle [6]. The result will foster the application of ML and AI algorithms by training them without the noise of anomalies or outliers.

The scientific novelty of this paper is the opportunity to apply this analysis to multiple sites simultaneously by approaching the problem with a clusterisation and a normalization applied in a portfolio of buildings. In literature, this kind of application was applied site by site exclusively. Another innovation of our study is the method used to define the best SAX representation of the dataset. The proposed methodology has been applied to a real-case buildings electrical consumption dataset with an hour resolution provided

by a telecommunication (TLC) service provider in Italy. More specifically, these data are a collection of electrical aggregate consumption of around 100 Central Offices (CO), managing the buildings of the national TLC network. Most of the electrical consumption of these buildings is related to TLC devices, with a constant consumption over the day, while the variable part of consumption is related to cooling systems of the environments to keep TLC equipment in a specific range of temperature.

The rest of the paper is organised as follows. Section II briefly reviews the literature on the topic. Section III presents the proposed algorithm. Section IV reports the experimental results obtained from analysing a real dataset composed of hourly measurements of the aggregate power data of TLC stations. Finally, Section V deals with the concluding remarks and future works.

II. Related Works

Defining a ground truth dataset to train machine learning algorithms in a robust way is not a trivial task. Moreover, this task is particularly complex for time series that shows variable behaviour over time and, in particular, what is expected in one period may be abnormal in another due to seasonal effects. Therefore, it is essential to adopt a procedure that can adapt to the dataset characteristics.

In the literature, there are several approaches taken to define the ground truth from a raw dataset. The main ones can be summarised into three categories: *i)* methods based exclusively on statistical principles, *ii)* clustering methods, and *iii)* methods based on dimensional reduction. The first approach is the simplest and is based on the definition of limits obtained through statistical analysis. A good example of such an approach is developed in [7]. While this approach leads to appreciable results, it can only be applied to dataset whose distribution can be assimilated to a Gaussian distribution. Moreover, this type of approach is more suited to detecting noise in the information rather than identifying real anomalies. These issues represent a major limitation to the use of this methodology.

Methods based on clustering through ML algorithms offer accurate dataset cleaning. In [8], hierarchical clustering was exploited to this purpose. The methodology allowed cleaning the dataset from anomalous behaviours and enhanced the training capabilities of an Auto-Regressive Integrated Moving Average (ARIMA) model. In [9], Liu et al. instead proposed the extraction of the ground truth from a raw dataset through the DBSCAN algorithm.

The third type is based on the dimensional reduction of the dataset. This approach highlights the salient aspects of the data and exclude redundancy and noise in the information. The different representation of the dataset is exploited to diversify anomalous behaviour from normal behaviour. In [10], Chicco et al. propose the use of Self-Organizing Map (SOM) to determine load profiles marked by anomalous consumption and exploit the results to improve the load forecast obtained by the application of a neural network. In [11], an application of Principal Component Analysis (PCA) reduction coupled with the use of unsupervised models provided the detection of anomalous patterns in an aggregated consumption dataset of the principal TLC service provider in Malaysia. In [12] the time series was grouped according to the boundary conditions. Finally, for each of the typical conditions, entropy definition was used to summarise the load profiles and identify whether they are anomalous or not. In Sial et al. [13] an ensemble of four different heuristics was used in order to extract different types of information from each methodology. This technique was applied on load profiles of different users within a student residence. However, the aforementioned methodologies transformed the dataset making it difficult for a system administrator to interpret these results and, subsequently, use them for energy management purposes. An alternative solution has been identified by Keogh et al. that proposed in [14] an innovative SAX approach that defines a dataset coding to reduce its size and discretise its variables still allowing the interpretability of the information contained in the dataset. This process unequivocally leads to a loss of some details but allows the dataset to be organised in a hierarchical manner by exploiting the information of the recurrence of a profile extracted through the discretization of the dataset. In fact, the SAX transformation allows the profiles of the dataset to be grouped according to the word through which they are encoded [15]. Subsequently, various researches were carried out in order to improve the encoding [16]–[20]. In particular, two different approaches have been proposed to improve this type of transformation. The first approach involves extracting more features from the dataset during the transformation reducing the loss of information related to coding. However, it complicates the use of the transformed dataset by requiring a clustering algorithms to reorganise it. Instead, the second approach optimises the coding itself, making it more flexible and adaptable to the dataset under examination. Moreover, another aspect of innovation introduced by this type of coding is the self-determination of the parameters that characterise SAX application. This aspect has been treated in [21]. However, its methodology is not applicable to a general scenario because used an objective function extrapolated by the context of knowledge that provided the best clusterization of the dataset.

The SAX coding has been successfully used also in the analysis of time series of electrical consumption of school buildings in [5], [20] that proposed an effective methodology for the treatment of this kind of data. However, no report has been found in the literature that deals with the definition of anomalous behaviour by exploiting the comparison between the behaviour of similar buildings through the use of SAX coding in order to extract the most significant load profile shapes of the population and exclude the anomalous profiles thus building a synthetic ground truth of consumption data.

This paper enhances the literature of synthetic ground truth generation by applying a normalization process that allows a fair comparison of the consumption of homogeneous COs of different sizes. This approach has not been much explored in the literature. Of all the articles examined, only [13] adopts a similar approach, albeit on a markedly different dataset. Defining anomalies based on the comparison of similar buildings allows a more robust load curve benchmark to be defined for the different boundary conditions to which the building is subjected. Afterwards, the SAX transformation is applied, exploiting different time intervals and energy intervals. The optimal combination results by applying a sensitivity method exploiting MIA index. The optimal combination, unlike the studies set out in [20] and in [5], is selected by using the Elbow method on the resulting MIA matrix of combinations. Finally, the Pareto principle was applied to determine whether

the resulting clusters (i.e. words) are considered as ground truth or anomalies. The proposed methodology is applied to a real TLC buildings dataset with different sizes. In particular, the dataset used consists of COs that manage the national TLC network. The analysis is carried out using their hourly electrical load pattern and information about the COs, such as their final use and geographical locations.

III. METHODOLOGY

The proposed methodology eliminates outliers and anomalies from a daily electricity consumption pattern dataset of industrial buildings to obtain synthetic ground truth. In particular, outliers and anomaly profiles have been identified by applying a SAX transformation to the proposed dataset. SAX allows to group the dataset according to the load profile shapes by exploiting a symbolic encoding (i.e. words). It is worth noting that through this methodology it is possible to identify the discordant load shapes through the recurrence of certain words within the dataset. The algorithm workflow is presented in Figure 1 and is based on three main steps: *i) Pre-processing*, *ii) Clustering*, and *iii) Post-Processing*.

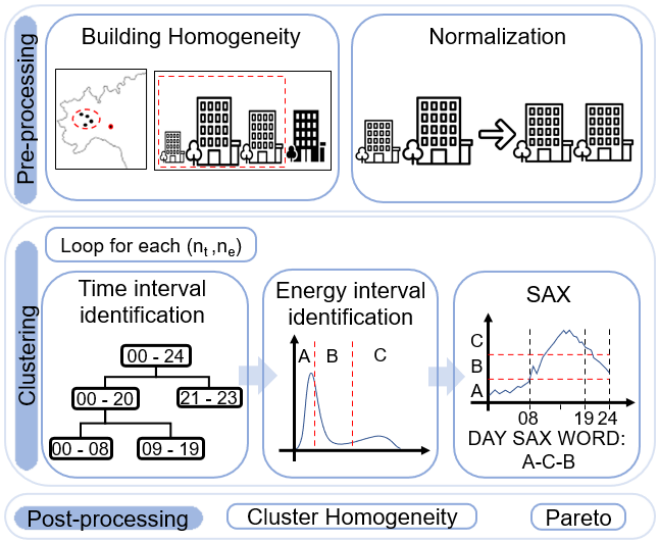

Fig. 1. The Synthetic Ground Truth Generation workflow

The *Pre-processing* step groups the homogeneous buildings as input for the following steps. The definition of homogeneity is obtained by two factors, namely *i)* the building's final use and *ii)* the geographic proximity. Once the homogeneous groups have been identified, the normalization process takes place to compare buildings featuring different sizes. In order to clearly define the normalization carried out, it is necessary to make some preliminary remarks on the components that define the load profiles of the TLC site dataset under analysis. Specifically, the electricity consumption can be decomposed into *i)* the energy spent to power the TLC components E_{TLC}, *ii)* the energy spent for the air conditioning systems $E_{CLC}(t)$, *iii)* the energy spent for Uninterrupted Power Supply (UPS) $E_{AUX}(t)$, and, finally, *iv*) the energy dissipated during the electric power transformations E_{DISS} (e.g. AC/DC). Therefore, the energy balance can be expressed as follows:

$$E_{TOT}(t) = E_{TLC} + E_{CLC}(t) + E_{AUX}(t) + E_{DISS}(t) \quad (1)$$

The E_{TLC} component is normally time-independent and is the only energy expense that corresponds to a positive economic balance [22]. Therefore, it is common to define TLC site efficiency indices based on this parameter. In this regard, the best indices are Power Usage Effectiveness (PUE) in Equation (2) and Utilisation Factor (UF) in Equation (3).

$$PUE(t) = \frac{E_{TOT}(t)}{E_{TLC}} \quad (2)$$

$$UF = \frac{E_{TLC}}{E_{min}} \quad (3)$$

While these metrics are the best descriptors of TLC site efficiency, there is no effective methodology in the literature for calculating E_{TLC} from aggregate consumption data. Therefore, the normalization has been carried out by estimating E_{TLC} as the E_{min}. This normalization method allows the definition of a new efficiency metric, so-called the Power Intensity Factor (PIF), which can be expressed as the product of PUE and UF, as follows:

$$PIF(t) = \frac{E_{TOT}(t)}{E_{min}} = \frac{E_{TOT}(t)}{E_{TLC}} \times \frac{E_{TLC}}{E_{min}} = PUE(t) \times UF \quad (4)$$

As the time series is composed of real data, it is necessary to find a method for evaluating E_{min} that is affected to a limited extent by the presence of outliers, which is why the E_{min} coefficient is calculated by evaluating the average of the minimum consumption values during the night hours (from 01:00 to 04:00) of January and February [23].

The *Clustering* step is composed of three main tasks: *i)* the Time Interval Identification, *ii)* the Energy Interval Identification, and, finally, *iii)* the SAX application. The Time Interval Identification aims to represent load daily profiles using a variable number of n_t time periods. This process is usually achieved by dividing the hours of the day into n_t equal time intervals. However, this method is not the most effective as the electrical load profiles of the facilities have periods in which the electrical load varies abruptly and periods in which consumption remains constant. During the periods of greatest load fluctuation, it is beneficial to reduce the width of the time windows to capture the shape of the load profile accurately. On the other hand, during the phases in which the electrical load remains stable, it is useless and counterproductive to apply time windows with small widths. An innovative solution to this problem is proposed in [20]. This process applies a CART to generate n_t time periods of variable amplitude by optimally defining their boundaries. The only independent variable supplied to the model is the time of day, while the target variable is the power required by the station. Therefore, the optimisation algorithm makes it possible to identify time windows across the leaf nodes that are characterised by the smallest variance in electricity consumption. Unlike the methodology proposed in [20], our methodology defines the best number of splits n_t downstream of the SAX transformation through a sensitivity analysis on all buildings pertaining to an homogeneous group. So, an exhaustive number of *time split case*, one for each n_t in range $[3, 7]$, are processed in the Energy Interval Identification task to identify the most sensitive configuration (n_t, n_e), with n_e representing the number of defined energy intervals.

The Energy Interval Identification process receives as input each *time split case* and calculates the mean value of the electric consumption pattern of all n_t time periods. Once

completed this process, all daily consumption patterns are reduced to n_t mean energy values. Afterwards, it calculates the Probability Density Function (PDF) of the mean values of the homogeneous group and generates different *energy split case* for each *time split case* by dividing for a variable number n_e of equally probable energy intervals. Also in this case, n_e varies in $[3, 7]$. The above-mentioned mean calculation allows to reduce the effect on the PDF of large time interval with constant consumption that could provoke unbalanced weight in the Energy Interval Identification task. The fair probability of the intervals is ensured by calculating the quantiles of the dataset population. This division generates balanced energy intervals and is essential to avoid periods of maximum consumption being unjustifiably labelled as anomalous.

The SAX transforms each daily load profile into words composed by n_t letters using a dictionary of n_e characters that represent the combination (n_t, n_e). Daily load profiles that is grouped in a particular word determine a cluster.

The clusters resulting from all (n_t, n_e) combinations tested feed the *Post-processing* step. Among the (n_t, n_e) combinations under test, the optimal combination is determined by comparing each combination MIA [24] in accordance with the so-called Elbow method by the Cluster Homogeneity task. MIA index is an indicator that expresses the goodness of the clustering case (n_t, n_e) under examination through the average of intra-cluster homogeneity index.The intra-cluster homogeneity index is calculated for each word identified during the SAX step. This index is equal to the daily average of the Euclidean distance of each profile from the cluster centroid (5).

$$WH_k = \frac{1}{N_{d^k}} \sum_{d^k}^{N_{d^k}} \left(\frac{1}{24} \sum_{t=1}^{24} |c_{k,t} - P_{d^k,t}|^2 \right)^{1/2} \quad (5)$$

MIA is evaluated as the average of the intra-cluster homogeneity values obtained for all words for a given combination (n_t, n_e).The MIA index is expressed by the following Equation:

$$MIA = \frac{1}{N_k} \sum_{k=1}^{N_k} WH_k \quad (6)$$

where:

- WH_k is the intra-cluster homogeneity for the k^{th} cluster;
- N_k is the number of clusters corresponding to the number of different words detected by the SAX;
- N_{d^k} is the number of days belonging to the cluster;
- $c_{k,t}$ and $P_{d^k,t}$ are respectively the hourly power values of the k^{th} cluster centroid and the daily load profile d^k belonging to the k cluster respectively.

Finally, the Pareto principle is applied to eliminate less significant clusters from the optimal combination (n_t, n_e) resulting by the Cluster Homogeneity task. The Pareto principle states that 20% of the causes are responsible for 80% of the effects. This principle was first set out by Pareto [6] and, nowadays, it is commonly adopted for quality analysis (Total Quality Management, Six Sigma, ISO9000) [25]. Following this principle, the Pareto block is employed to assess whether a profile is significant or anomalous. In particular, the dataset is ordered according to the recurrence of the SAX encoding, i.e. according to the magnitude of each cluster. The 20% of the clusters defined by the most recurrent words is expected to host about 80% of the dataset. This recurrent patterns are therefore considered to be the significant load profiles for the buildings under examination. On the contrary, the remaining part of the dataset shall be distributed in the 80% of the minor clusters. This part of the dataset is very sparse and can therefore be assumed to contain anomalous observations. These infrequent profiles shall be marked as outliers and removed from the dataset, as the exploitation of these values for training ML tools is may worsen the models accuracy.

IV. Experimental Results

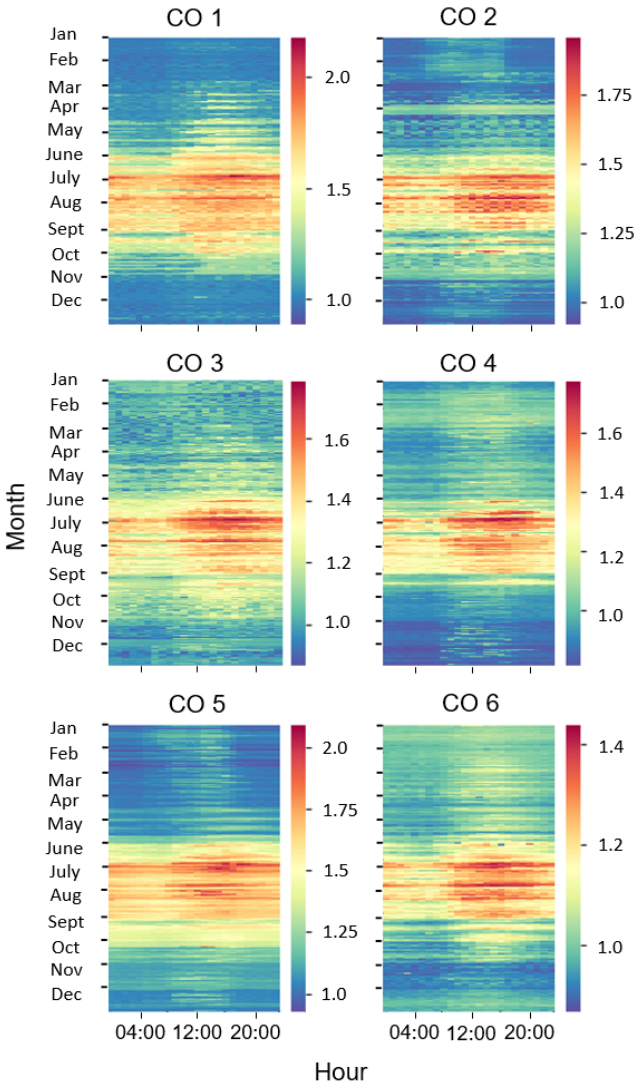

Fig. 2. Carpet plot representation of the selected homogeneous group of six Central Offices, the colour scale shows for each CO the PIF values

This section presents the results obtained by applying the proposed methodology to a dataset consisting of around 100 COs from a TLC provider in Italy. Through the process of Building Homogeneity, several building groups were identified. The methodology was applied individually to each identified group. For the sake of clarity, the experimental results of a group composed of 6 COs is reported to assess the proposed methodology. In Figure 2, the hourly load profiles of the selected building group are presented in the carpet plots after the application of the normalization process. These

graphs provide a concise representation of the entire dataset under examination, highlighting typical seasonal trends like the effect of cooling loads in summer days that rises the load consumption patterns during central hours of the day.

Once the Pre-processing step has been completed, the Clustering process is applied to the resulting load profiles time series. As previously introduced in Section III, this step initially involves testing different time and energy intervals of the SAX encoding in order to identify the optimal combination (n_t, n_e). The Time Interval and Energy Interval Identification tasks are so executed to identify the different combination of (n_t, n_e). Then, the SAX task applies to each combination the feature reduction generating a set of cluster defined by their specific words composed by n_t character on a n_e dictionary.

The results of Clustering step are then sent to the Post-processing step. The Cluster Homogeneity task composes a matrix of MIA values for each of the combination n_t in $[3, 7]$ and n_e in $[3, 7]$. The resulting matrix has been depicted through the three-dimensional histogram shown in Figure 3. This representation makes it easy to identify the selected Elbow point of the MIA matrix. In particular, $(4, 4)$ is identified as the best time and energy interval parameters configuration from the Cluster Homogeneity step.

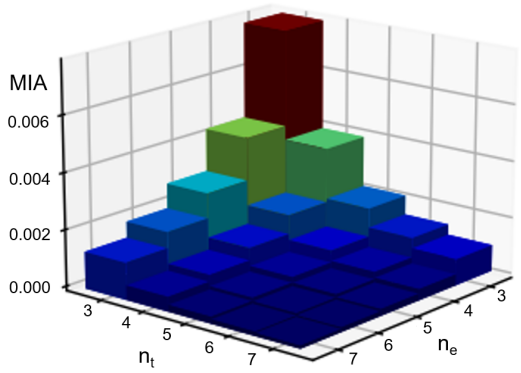

Fig. 3. Graphical representation of the MIA matrix considering the different combinations (n_t, n_e)

Once the best configuration is selected, the results of the SAX algorithm concerning the $(4, 4)$ configuration is retrieved to feed the Pareto task. The time interval identification of the $(4, 4)$ combination identifies with the CART four time intervals ($[00{:}00 - 09{:}00], [10{:}00 - 11{:}00], [12{:}00 - 19{:}00], [20{:}00 - 23{:}00]$). The resulting segmentation has a good correspondence with the typical phases that characterise the consumption pattern of a CO. The first time interval corresponds to the low consumption range of the CO. The next window $[10{:}00 - 11{:}00]$ identifies the period of sudden load variations, which is commonly called system rump up. The central time interval $[12{:}00 - 19{:}00]$ represents the period of maximum consumption of the CO. Finally, the window $[20{:}00 - 23{:}00]$ represents the period when consumption settles down to low levels. At the end of this phase, the load profiles are reduced in size by calculating the average PIF for each identified time window. In this way, each daily observation will no longer be composed of 24 variables but only of 4 mean energy consumption values.

The energy interval identification of the $(4, 4)$ combination receives the dataset from the previous step to identify four equiprobable PIF energy intervals. As reported in Figure4, the distribution of the PIF population is markedly unbalanced towards low values. Therefore, the energy intervals will be denser for low values and sparser for high values of PIF. This aspect represents a weak limitation of the SAX transformation, which is not adequately detailing high consumption. However, the equal probability of occurrence of each power interval is an essential characteristic for the correct application of the methodology.

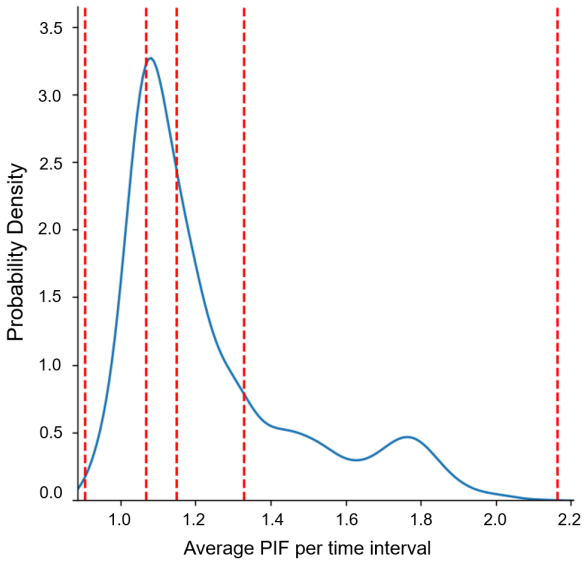

Fig. 4. Graphical representation of probability density function (PDF) of the PIF population and its four equiprobable energy intervals defined by the Energy Interval Identification task

Once the time interval and energy interval calculations have been concluded, it is possible to describe each day through a string composed of n_t letters, each of which can have n_e characters. Therefore, this type of representation allows a number of possible combinations equal to $N_{comb} = n_t^{n_e}$, which in the case of $(4, 4)$ SAX configuration the number of words that can be generated is 256. However, only 65 words result from the entire dataset.

In order to clearly define the limit between motifs and discord, the Pareto principle was exploited by defining as motifs only the 80% of the overall population, selecting the largest clusters defined by the Clustering step and applying the combination $(4, 4)$. In Figure 5, the Pareto diagram shows on the x-axis the daily encodings for each of the 65 resulting words ordered according to the recurrence of each code within the dataset. The information on the recurrence of each word is reported in a quantitative form on the left axis. The right axis instead shows the evolution of the cumulative percentage of the selected words. Through this representation, it is possible to verify the applicability of the Pareto principle to the selected group. As shown in Figure 5, the 13 most recurrent symbols, that is around 20% of the 65 identified words, represents around 84% of the dataset. Moreover, the remaining part of the dataset (16%) is represented by 52 different symbols. The dispersion of this part of the dataset is an obstacle to a proper learning of ML algorithms with a consequent deterioration in their performance. Indeed, the exclusion of the these outliers from the dataset determines a consistent improvement of the homogeneity index of the dataset, from a value equal to 0.170 for the original dataset to

0.194 for the filtered one. For the sake of readability, only the 21 most frequently occurring SAX words have been included in the plot in Figure 5.

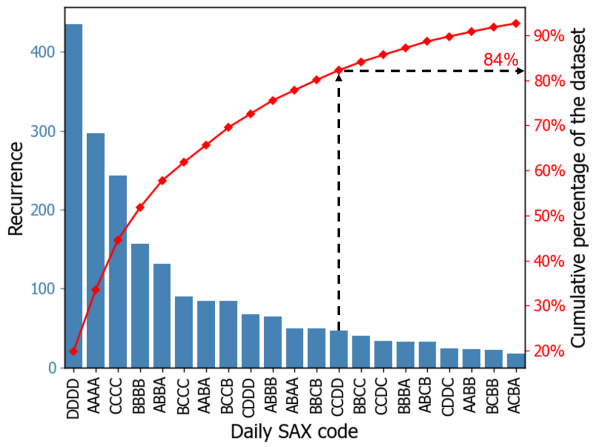

Fig. 5. Pareto diagram of the resulting $(4, 4)$ Clustering step.

V. Conclusion

This paper presents a novel methodology to define a synthetic ground truth from load profile time series of industrial sites, eliminating daily anomalies and outliers. The proposed SAX evolution applies a proper pre-processing to identify discord load profiles on a set of buildings rather than considering individually each building. This innovative solution allows highlighting not only statistical anomalies but also anomalies linked to the non-optimal management of industrial sites of a homogeneous group. By ordering the resulting SAX words from the largest to the lowest recurrence and applying the Pareto principle, the synthetic ground truth is generated as the first 80% of the dataset. The selected 80% represents the 20% of the resulting SAX words, confirming the Pareto principle. The load profiles outside this selection are considered anomalies and discarded. The process described is autonomous and does not require parameter imputation by the analyst. However, it is important to emphasise the importance of a system expert to evaluate and validate the results of the algorithm. Such validation, although costly in terms of user time, is certainly facilitated by SAX encoding and the graphical representation of the entire dataset by means of carpet plots. In future works, the methodology will be validated on a public dataset in which consumption anomalies have been previously labelled. In addition, different ML and AI models for anomaly detection will be tested by training them on the overall dataset or on the synthetic ground truth dataset and comparing the performance obtained according to the different data used for training. The expected result is that anomaly detection will benefit from our filtering process and will ensure better anomaly detection results.

References

[1] IEA. Key World Energy Statistics 2021. (August 2021). [Online]. Available: https://www.iea.org/reports/electricity-information-overview/electricity-consumption

[2] Z. Zhongming, Z. Wangqiang, L. Wei *et al.*, "World energy balances overview (2020 edition)," 2020.

[3] S. Katipamula, R. M. Underhill, N. Fernandez, W. Kim, R. G. Lutes, and D. Taasevigen, "Prevalence of typical operational problems and energy savings opportunities in us commercial buildings," *Energy and Buildings*, vol. 253, p. 111544, 2021.

[4] M. Gaur, S. Makonin, I. V. Bajić, and A. Majumdar, "Performance evaluation of techniques for identifying abnormal energy consumption in buildings," *IEEE Access*, vol. 7, pp. 62 721–62 733, 2019.

[5] C. Miller, Z. Nagy, and A. Schlueter, "Automated daily pattern filtering of measured building performance data," *Automation in Construction*, vol. 49, pp. 1–17, 2015.

[6] R. Sanders, "The pareto principle: its use and abuse," *Journal of Services Marketing*, 1987.

[7] S. Maleki, S. Maleki, and N. R. Jennings, "Unsupervised anomaly detection with lstm autoencoders using statistical data-filtering," *Applied Soft Computing*, vol. 108, p. 107443, 2021.

[8] J.-S. Chou and A. S. Telaga, "Real-time detection of anomalous power consumption," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 400–411, 2014.

[9] X. Liu, Y. Ding, H. Tang, and F. Xiao, "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data," *Energy and Buildings*, vol. 231, p. 110601, 2021.

[10] G. Chicco, R. Napoli, and F. Piglione, "Load pattern clustering for short-term load forecasting of anomalous days," in *2001 IEEE Porto Power Tech Proceedings (Cat. No. 01EX502)*, vol. 2. IEEE, 2001, pp. 6–pp.

[11] M. Jesmeen, J. Hossen, and A. B. A. Aziz, "Unsupervised anomaly detection for energy consumption in time series using clustering approach," *Emerging Science Journal*, vol. 5, no. 6, pp. 840–854, 2021.

[12] X. Zhou, T. Yang, L. Liang, X. Zi, J. Yan, and D. Pan, "Anomaly detection method of daily energy consumption patterns for central air conditioning systems," *Journal of Building Engineering*, vol. 38, p. 102179, 2021.

[13] A. Sial, A. Singh, and A. Mahanti, "Detecting anomalous energy consumption using contextual analysis of smart meter data," *Wireless Networks*, vol. 27, no. 6, pp. 4275–4292, 2021.

[14] P. Patel, E. Keogh, J. Lin, and S. Lonardi, "Mining motifs in massive time series databases," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 2002, pp. 370–377.

[15] E. Keogh, J. Lin, and A. Fu, "Hot sax: Efficiently finding the most unusual time series subsequence," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*. Ieee, 2005, pp. 8–pp.

[16] S. Yang, Y. Wang, and J. Zhang, "A similarity measure for time series based on symbolic aggregate approximation and trend feature," in *2020 39th Chinese Control Conference (CCC)*, 2020, pp. 6386–6390.

[17] K. Zhang, Y. Li, Y. Chai, and L. Huang, "Trend-based symbolic aggregate approximation for time series representation," in *2018 Chinese Control And Decision Conference (CCDC)*, 2018, pp. 2234–2240.

[18] Y. Zhang, L. Duan, and M. Duan, "A new feature extraction approach using improved symbolic aggregate approximation for machinery intelligent diagnosis," *Measurement*, vol. 133, pp. 468–478, 2019.

[19] N. D. Pham, Q. L. Le, and T. K. Dang, "Two novel adaptive symbolic representations for similarity search in time series databases," in *2010 12th International Asia-Pacific Web Conference*, 2010, pp. 181–187.

[20] A. Capozzoli, M. S. Piscitelli, S. Brandi, D. Grassi, and G. Chicco, "Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings," *Energy*, vol. 157, pp. 336–352, 2018.

[21] M. S. Gallimore, C. M. Bingham, and M. J. W. Riley, "Self-organising symbolic aggregate approximation for real-time fault detection and diagnosis in transient dynamic systems," in *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2017, pp. 000 043–000 048.

[22] M. Sorrentino, M. Bruno, A. Trifirò, and G. Rizzo, "An innovative energy efficiency metric for data analytics and diagnostics in telecommunication applications," *Applied Energy*, vol. 242, no. March, pp. 1539–1548, 2019. [Online]. Available: https://doi.org/10.1016/j.apenergy.2019.03.173

[23] S. Eiraudo, L. Barbierato, R. Giannantonio, A. Porta, A. Lanzini, R. Borchiellini, E. Macii, E. Patti, and L. Bottaccioli, "Load profiles clustering and knowledge extraction to assess actual usage of telecommunication sites," in *2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*. IEEE, 2021, pp. 1–6.

[24] S. A. Notari, G. Chicco, and F. Piglione, "Data size reduction with symbolic aggregate approximation for electrical load pattern grouping," *IET Generation, Transmission & Distribution*, vol. 7, no. 2, pp. 108–117, 2013.

[25] L. Wilkinson, "Revising the pareto chart," *The American Statistician*, vol. 60, no. 4, pp. 332–334, 2006.