Empirical Exploration of Energy Estimation Metrics in Load Disaggregation

João Góis
Interactive Technologies Institute / LARSyS
Instituto Superior Técnico, University of Lisbon
Email: joao.gois@iti.larsys.pt

Lucas Pereira
Interactive Technologies Institute / LARSyS
Instituto Superior Técnico, University of Lisbon
Email: lucas.pereira@iti.larsys.pt

***Abstract*—Improving energy consumption feedback to end-users is crucial for energy optimization in buildings, thus reducing the impacts of energy waste on climate change. Non-intrusive Load Monitoring became a relevant approach for energy estimation in buildings. However, a more robust knowledge of the performance metrics used for evaluating the disaggregation algorithms is necessary for more accurate interpretation and conclusions. This work empirically analyses how the relationships between performance metrics vary across different appliance types and how they influence metric selection. Our results demonstrate that the kind of appliance disaggregated significantly influences metric relationships, providing critical insights for selecting an optimal set of metrics.**

I. Introduction

In the last three decades, there has been a significant research effort to mitigate global climate change. In particular, the energy consumption in buildings accounted for 25% of global CO2 emissions in 2019 [1]. Hence, energy efficiency became a key target for climate action, as stated clearly in the United Nations' Sustainable Development Goals 7, 11, and 13 [2]. The advent of smart meters [3] and of Non-Intrusive Load Monitoring (NILM) [4] brought a low-cost manner of monitoring appliance energy consumption by estimating individual appliance consumption from the total energy (aggregate) signal, using only one sensor/smart meter. The costs with infrastructure are reduced in comparison with other monitoring methods, facilitating its implementation in real-world applications [5].

Two NILM modeling approaches emerged: event-based and eventless methods [6]. Event-based methods focus on classifying appliance transitions, while eventless methods optimize overall system estimation through regression tasks. Over time, eventless methods have been preferred due to their label-free nature and many available datasets. To evaluate eventless approaches, metrics that measure errors between estimated and actual energy consumption — known as energy estimation (EE) metrics — are considered [7], [8].

Despite the number of EE metrics utilized in NILM, it is still challenging to define a consistent and widely accepted set of metrics [6], [9], [10]. Traditional EE metrics, like RMSE, are unsuitable for comparing performance across different appliance types [11], while there is a lack of knowledge on how recent EE metrics handle varying appliance usage patterns and power. A deep understanding of the metrics is essential for accurate NILM evaluation [6]. For this purpose, studying metric relationships across different appliance types could be critical to help select a suitable set of metrics, which was not sufficiently explored in previous studies [12].

This work conducts an empirical study to assess to what extent metric relationships vary across different appliance types in the context of energy estimation. To this end, the correlations between the metrics are assessed by considering hierarchical clustering analysis for each appliance using real-world NILM data. The remainder of this work is structured as follows: Section II provides background and related work on the topic; Section III provides the setup and results of the empirical experiment, and Section IV provides the main conclusions, limitations, and implications of this work.

II. Related Work

The selection of metrics for performance evaluation is crucial for assessing the algorithms' accuracy and understanding appliance energy consumption. Some research works have been dedicated to studying metric relationships empirically or analytically.

Analytical approaches typically focused on the theoretical properties of metrics across different energy estimation scenarios, such as in Mayhorn et al. [12]. Normalized EE metrics like the match rate, energy accuracy, and energy error were suggested for evaluating the algorithms, in contrast with other traditional metrics, like RMSE. Nonetheless, a careful examination of these metrics is thus recommended when dealing with different appliance usage patterns across datasets.

Empirical approaches typically involve running algorithms across various datasets. Pereira et al. conducted this kind of approach to inspect the metric relationships when applied to classification algorithms in event-based NILM [7]. The correlations between performance metrics were analyzed via hierarchical clustering analysis. The results showed a low correlation between metrics based on the confusion matrix, as is also the case in other machine learning domains. Additionally, the authors suggested that probabilistic (energy estimation) metrics should be assessed in future work to provide more insights into NILM performance, leveraging the distance between estimated and correct values.

Furthermore, Pereira et al. also focused on defining a consistent set of metrics for evaluating event detection algorithms

[10]. An empirical exploration was conducted to analyze metric relationships, showing significant changes compared to other machine learning domains. The unbalanced nature of the event-detection problem was identified as a potential cause for such results and the authors recommend that new metrics consider the power levels of datasets.

In this work, an empirical analysis of metric relationships in energy estimation NILM and their influence on metric selection is conducted. Different appliance types were considered, offering insights that can guide the selection of metrics for more accurate performance evaluation.

While previous works [7], [10] have comprehensively explored metric relationships in event-based NILM, this study is the first to empirically analyze energy estimation metrics in NILM, addressing a key gap in the literature. The following empirical study builds on this body of work by using correlation and hierarchical clustering analysis to examine how EE metrics behave across different appliance types.

III. Empirical Study

A. Dataset

The study considers the REFIT [13] dataset, which contains household consumption data at both aggregate and individual levels over approximately two years. The analysis specifically considers eight appliances (dishwasher, fridge freezer, fridge, freezer, kettle, microwave, television, washing machine) with different power levels and usage patterns to better compare metric relationships across the households. For illustration, the time resolution of the data is equal to one sample per minute, which enables a suitable trade-off between data frequency and computational complexity of experiments. According to [14], the classification of appliance types in this work is the following: washing machines, dishwashers, and microwaves have multiple operation states (type II) and are user-dependent; fridge freezers, fridges and freezers are also type II, but user-independent; kettles are on-off appliances (type I) and user-dependent; and finally, televisions are user-dependent and are considered type IV appliances since their receivers are permanently active, consuming energy.

B. Algorithms

The disaggregation experiments were carried out using 4 baseline algorithms in NILM, which are well-known in the field, with widely recognized performance.

- Edge detection (ED) — Proposed by Hart [4], this baseline NILM algorithm detects edges (transient or steady states) in a signal by calculating power differences between adjacent timestamps. Although ED is unsupervised, the training data is used for matching edges to appliances for better accuracy.
- Combinatorial Optimization (CO) — Proposed by Hart [15], this algorithm finds the optimal combination of appliance states that add up to the observed total power for each time slice. Each time step is optimized independently.
- Mean — Predicts appliance usage as the mean computed on the training data. This algorithm is a reliable baseline against more complex algorithms and can be especially useful for disaggregating sparse appliances.
- Exact Factorial Hidden Markov Model (FHMME) — Models each appliance power demand and states as the observed and hidden components of a Hidden Markov Model (HMM), respectively. Then, the individual HMMs are combined to jointly infer appliance states from the total power signal [16], [17].

For running the algorithms, the training set was defined from June 2014 to December 2014, and the test set from January 2015 to June 2015. The previous algorithms are implemented in NILMTK [18], [19], which facilitated the disaggregation experiments across REFIT households (summarized in Table I).

C. Performance Metrics

Since eventless approaches are used for modeling the NILM problem, EE metrics are utilized for assessing the quality of disaggregation, by measuring the error between actual and estimated energy. A total of 15 metrics are considered, ranging from traditional multidisciplinary metrics, such as RMSE and MAE, to recent metrics proposed specifically for NILM, like the MR and FEE. Table II provides a brief description of each metric, *vide* [6] for further detail.

Since metrics like ABSE, AE, MAE, RMSE, SDE, SEM, and ETEA are non-normalized, performance comparison must be conducted cautiously when using these metrics [11], [20]. In particular, for AE and ETEA, lower scores do not necessarily mean a good performance; it is only certain that the amounts of overestimation and underestimation are similar.

The non-normalized metrics are confronted with normalized metrics, which are more suitable for comparing performance across different appliance types, such as CVRMSD, EE, EA, MR, FEE, R2, PSDE, and DEV. The CVRMSD is a normalization of RMSE by the mean [21], [22]. According to [11], the total appliance energy consumption should be used as a normalization constant, like the EE, EA, FEE, R2, PSDE, DEV, and MR metrics. This would prevent the possible outlier problem of using the mean. Notice that EA results from re-scaling EE to [0, 1], similarly for PSDE and R2. The DEV metric should be interpreted carefully, as it derives from ETEA. For the MR, the normalization constant is greater or equal to the total energy consumption. The MR, in particular, provides the overlapping rate between true and estimated energy, which combined with being normalized between 0 and 1, made this metric recommended for evaluating the performance in NILM problems [12].

D. Experiment Setup

To explore the hypothesis that the type of disaggregated appliance influences the choice of the metrics, the following steps were considered:

TABLE I
SUMMARY OF DISAGGREGATION EXPERIMENTS FOR EACH APPLIANCE AND ALGORITHM BEFORE THE CALCULATION OF THE METRICS.

	Dishwasher	**Fridge freezer**	**Fridge**	**Freezer**	**Kettle**	**Microwave**	**Television**	**Washing machine**	**Total**
CO	15	14	7	11	14	16	19	19	115
FHMME	15	16	7	11	14	16	19	19	117
ED	15	14	7	11	14	16	19	19	115
Mean	15	14	7	11	14	16	19	19	115
Total	60	58	28	44	56	64	76	76	462

1) For each appliance i, disaggregation algorithm $j = 1, \cdots, 4$ is run across the houses and metrics $j_1, \cdots, j_{15}$ are calculated for each case;
2) Calculate the pairwise correlations between $j_1, \cdots, j_{15}$ using rank (Spearman) correlation [23], leading to correlation matrices $C^{\star}_{i,j}$ in each case;
3) Average the correlation matrices $C^{\star}_{i,j}$, $j = 1, \cdots, 4$ to obtain an average matrix C_i;
4) Repeat for $i + 1$.

The Spearman method was considered instead of the Pearson (linear) method because it also considers potential nonlinear relationships between the metrics. The choice for averaging correlation matrices for different algorithms was based on the fact that rank correlations between metrics are likely to vary across the algorithms. The same procedure was considered to calculate a matrix of average correlations when considering the experiments of all the appliances at once (referred to as the total set of appliances). This will enable the comparison of correlations between the metrics in two scenarios: appliance level and total set of appliances.

By inspecting the performance metric scores, the AE metric could pose a problem when calculating the rank correlations, as it takes either negative or positive values when there is a higher amount of overestimation or underestimation on average, respectively. The optimal value for the AE is zero. To ensure the ranks are obtained appropriately, the absolute values of AE were considered.

For the hierarchical clustering analysis, the clustering method's input should indicate the dissimilarity between objects, i.e., it should reflect the distances between each pair of metrics. For this purpose, the correlation matrices C_i were preprocessed by applying to each entry (l, m) the dissimilarity function defined by

$$D_{i,lm} = 1 - |C_{i,lm}| \tag{1}$$

This function was proposed for graphical representations using dendrograms [24] and ensures that pairs with "stronger" correlation are ordered correctly from the bottom ($|C_{i,lm}| = 1$) to the top ($|C_{i,lm}| = 0$). Also, negative and positive correlations with the same absolute value have equal dissimilarity. Then, the dissimilarity matrices D_i are transformed into lower triangular matrices, $D_{i,low}$ for applying hierarchical clustering. The metrics are grouped into clusters using the Ward method as a linkage function, which finds the clustering that minimizes variance within each cluster [25].

The formation of clusters can be visualized through the dendrograms, enabling one to understand how metrics cluster together. In the dendrograms for each appliance and the total set of appliances, the cut-off line to decide the number of metric clusters is chosen considering the maximization of the silhouette score. Fig. 1 shows the correlation matrix and the dendrogram obtained from the correlations verified for the total set of appliances.

Finally, the rand index (RI) [26] is considered to analyze the similarity between the clusterings of metrics obtained for each appliance and the total set of appliances. RI computes the closeness between two clusterings by comparing the number of agreeing pairs of metrics (never clustered together/clustered in both clusterings) to all possible pairs, given by

$$\text{RI} = (\text{number of agreeing pairs})/(\text{number of pairs}) \tag{2}$$

RI varies between 0 and 1, where a higher value indicates more closeness. Notice that the total number of pairs includes the disagreeing pairs, i.e., metrics switch between clustered and not clustered together across clusterings.

After computing the rand indices, it is investigated whether metric relationships vary depending on the appliance type. For that purpose, the appliances are grouped according to the rand indices to assess the potential relationships between the appliances.

E. Results and Discussion

Following the procedure described in the previous subsection, the dendrograms resulting from hierarchical clustering were obtained for the individual appliances (Fig. 1 already showed the dendrogram for the total set of appliances). Then, the number of clusters for each case is determined after consulting the silhouette scores.

For the total set of appliances, a higher silhouette score is observed if 7 clusters are selected. As for each appliance, the number of selected clusters varies significantly, from 2 to 9. Fig. 2) shows the clusterings for each case in more detail. The rand indices calculated between the appliances' clusterings suggest similar metric relationships for the television, dishwasher, washing machine, microwave, fridge freezer, and freezer. The visual inspection of Fig. 2 suggests that these appliances are more similar to the total set of appliances, which can be confirmed by calculating the rand indices.

In contrast, the kettle and fridge are suggested to have only 2 clusters, being significantly different from the other appliances. Additionally, the cutoff distance in the respective dendrograms

TABLE II
PERFORMANCE METRICS UTILIZED FOR EVALUATION.

Metric	Formula	Description
AE	$\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)$	The Average Error indicates if the estimated energy is, on average, over or underestimated. A positive (negative) score implies a higher amount of over(under)-estimation. When the score tends to zero, the amounts of over and underestimation are similar.
ABSE	$\sum_{t=1}^{T}\lvert\hat{y}_t - y_t\rvert$	The Absolute Error indicates about the total amount of error, whether it is over or underestimation.
MAE	$\frac{1}{T}\sum_{t=1}^{T}\lvert\hat{y}_t - y_t\rvert$	The Mean Absolute Error gives the average error on estimated energy, whether it is over or underestimation.
RMSE	$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}$	The Root Mean Squared Error or deviation is the standard deviation of the energy estimation errors. It informs on the degree of dispersion of the errors with respect to actual values. RMSE/RMSD is presented in the same unit as the data, making it an intuitive metric.
CVRMSD	$\frac{\text{RMSE}}{\bar{y}}$	The Coefficient of Variation consists of normalizing the RMSE by the mean of the actual values. The lower scores indicate less residual variance, and vice-versa.
EE	$\frac{\sum_{t=1}^{T}\lvert\hat{y}_t - y_t\rvert}{\sum_{t=1}^{T} y_t}$	The Energy Error is a normalized metric consisting of the ratio of total absolute errors and the total amount of actual energy.
EA	$e^{-\alpha \text{EE}}$	The Energy Accuracy is a normalized metric that standardizes the EE metric between 0 and 1. Mayhorn et al. [12] suggested using $\alpha = 1.4$.
MR	$\frac{\sum_{t=1}^{T}\min\{\hat{y}_t, y_t\}}{\sum_{t=1}^{T}\max\{\hat{y}_t, y_t\}}$	The Match Rate is a normalized metric that evaluates the overlapping rate of the true and estimated energy, varying between 0 (weak match) and 1 (strong match).
FEE	$\frac{\sum_{t=1}^{T}\hat{y}_t}{\sum_{t=1}^{T} y_t}$	The Fraction of Energy Explained is a normalized metric that confronts the total estimated energy to the total actual energy.
R^2	$1 - \sum_{t=1}^{T}\frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2}$	The R-squared is a statistical measure between 0 and 1 of the proximity between the estimated and actual data. The higher the R^2, the more the estimates are in line with the actual data.
PSDE	$1 - \sqrt{1 - R^2}$	The Percentage of Standard Deviation Explained is the percentage where the standard deviation of errors is lower than the standard deviation of the actual data. The PSDE can be more intuitive than R^2 since it is presented in the same units as the actual data.
SDE	$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\Delta y_t - \overline{\Delta y})^2}$	The Standard Deviation of the Error indicates the degree of dispersion of the error around the average error estimate. A larger SDE implies a wider dispersion of the estimated values, while a smaller SDE implies tighter distributions.
SEM	$\frac{\sigma}{\sqrt{T}}$	The Standard Error of the Mean indicates how different the estimated energy mean is likely to be from the actual sample mean. SEM increases for large errors, and vice-versa.
ETEA	$\lvert\sum_{t=1}^{T}\hat{y}_t - \sum_{t=1}^{T} y_t\rvert$	The Error in Total Assigned Energy quantitatively measures the total difference between all estimated and all actual energy.
Dev	$\frac{ETEA}{\sum_{t=1}^{T} y_t}$	The Deviation compares the difference between all estimated and all actual energy to the total actual energy.

would need to be reduced significantly to obtain clusterings for the kettle and fridge similar to those of the first group appliances, at the expense of reducing the silhouette score.

The fact that fridge and kettle have similar metric relationships was unexpected due to their different usage patterns. The fridge is a user-independent appliance that operates in recurrent cycles like the freezer and fridge freezer, while the kettle is associated with the consumer's routines, like the television and washing machine. The divergence between the kettle and the other user-dependent appliances may be due to the distinct power consumption pattern of the kettle, which consists of short, intense bursts of energy consumption, reflecting sporadic use. The fact that the fridge is not similar to other user-dependent appliances is also no surprise. However, the divergence between the fridge and the fridge freezer, and fridge and freezer was unexpected in this context. One potential reason is the differences in operation modes and usage patterns (open/close). The fridge freezer combines the operation modes of both a fridge and a freezer and is generally opened and closed more frequently than the fridge, leading to higher energy consumption. The freezer is generally opened and closed less frequently than the fridge, thus leading to less energy consumption [27]. Further experiments with other algorithms and datasets would be relevant for checking whether the results from this work hold.

In the first group of appliances, there are five type II and one type IV appliances, while in the second group, there are one type II and one type I appliances. Hence, there is evidence that the type of disaggregated appliance leads to different metric relationships, proving the initial hypothesis. The previous results also demonstrate the importance of analyzing metric relationships for each appliance individually, since the similarities and dissimilarities between appliances would probably go unnoticed if only the total set of appliances was considered.

To find a set of metrics to compare the performance between different appliances in this study, the metrics consistently paired for most appliances seem to have a relatively stable association and could be potential candidates for evaluation. The RMSE and SDE are clustered together for all appli-

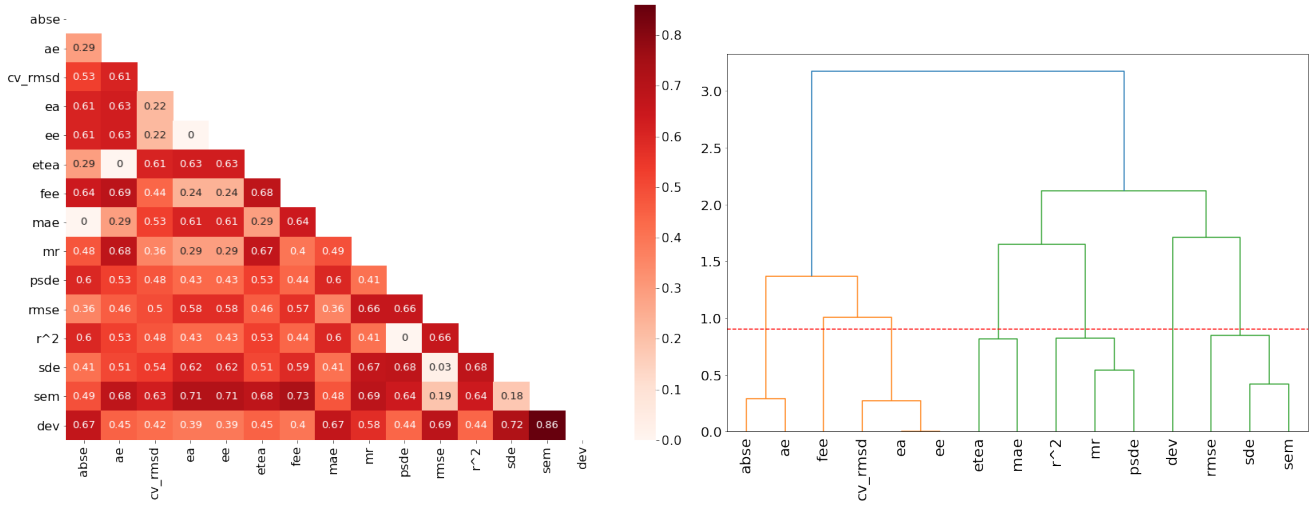

Fig. 1. Spearman correlations between metrics for the total set of appliances (left), and the dendrogram that indicates cluster partition of metrics along the cutoff distances (right). The partition with a higher silhouette score is crossed with a red line.

ances, which agrees with both metrics being based on the deviation. In the case of the ABSE and AE, the concepts are relatively similar. The FEE, CVRMSD, EA and EE are clustered together for all appliances, except the dishwasher. Notice that EA results from the standardization of EE, thus both metrics should be naturally clustered together. Finally, the MR, PSDE, and R2 are also clustered together for most appliances; in particular, the relationship between PSDE and R2 is analogous to that between the EA and EE. Notice that the metrics in (FEE, CVRMSD, EA, EE) or (MR, PSDE, R2) are normalized, which ensures consistent scaling and eliminates the influence of differing energy consumption levels when comparing performance across the appliances. Hence, one metric from each side could be chosen for this purpose. From the analytical analysis conducted in Mayhorn et al. [12], the MR, EA and EE have already been indicated as suitable candidates for performance evaluation, which seems to be aligned with the empirical analysis from this study, where such metrics were grouped in clusters that do not contain the RMSE for all appliances. Also, the fact that MR and EA range from 0 to 1 indicates that these two metrics are well-suited for providing reliable performance interpretation and comparison.

IV. Conclusion

This study addresses the gap in understanding how EE metric relationships are influenced by the type of disaggregated appliance, a topic insufficiently explored in previous NILM research. From the empirical analysis, there was evidence that the metric relationships varied across different appliance types, which seems to be related to the distinct power levels and usage patterns. The unexpected similarity between the fridge and kettle in metric relationships suggests that further investigation is needed to understand the underlying causes of this output. Similarly, the divergences between the fridge and fridge freezer, fridge and freezer, could be further explored.

It should also be noted that by conducting this analysis for datasets with varying appliance usage patterns and sample rates, the clusterings of metrics may vary, leading to the selection of different metrics. Nevertheless, the findings from this paper provide a foundation to compare performance across appliance types, with direct implications for the design of more accurate and adaptable NILM systems. By selecting the metrics that consistently correlate across different appliance types, taking into account their properties, energy providers can deploy NILM systems that provide more accurate feedback to users, leading to more informed energy-saving practices.

Future research could extend the analysis conducted in this work by exploring a wider range of appliances and datasets, as well as running more advanced NILM algorithms. Additionally, the analysis of correlations between metrics was based on rank correlations, which do not directly reflect the quality of disaggregation outputs; therefore, the exploration of a method for that purpose could be a suitable addition to this work. The previous extensions would enable the validation of the obtained findings and the improvement of our understanding of metric relationships.

Acknowledgment

This work received funding from the Portuguese Foundation for Science and Technology (FCT) under projects 10.54499/LA/P/0083/2020; 10.54499/UIDP/50009/2020 & 10.54499/UIDB/50009/2020, and grants DFA/BD/8075/2020 (J.G.) and CEECIND/01179/2017 (L.P.).

References

[1] https://www.iea.org/topics/climate-change, accessed: 2024-06-29.

[2] U. N. D. of Economic and S. Affairs, *The Sustainable Development Goals Report 2019*, 2019th ed. United Nations, 2019. [Online]. Available: https://www.un-ilibrary.org/content/books/9789210478878

[3] B. Völker, A. Reinhardt, A. Faustine, and L. Pereira, "Watt's up at Home? Smart Meter Data Analytics from a Consumer-Centric Perspective," *Energies*, vol. 14, no. 3, p. 719, Jan. 2021.

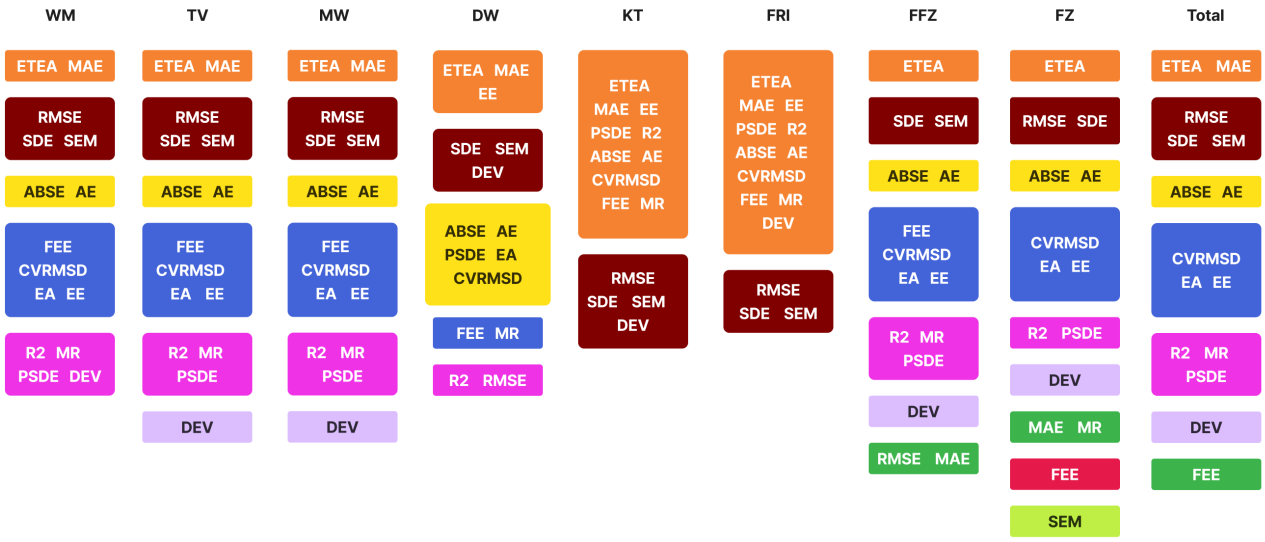

Fig. 2. Selected cluster partition of the metrics with a higher silhouette score for each appliance and total set of appliances (DW: Dishwasher, FFZ: Fridge Freezer, FRI: Fridge, FZ: Freezer, KT: Kettle, MW: Microwave, TV: Television, WM: Washing Machine).

[4] G. W. Hart, "Prototype nonintrusive appliance load monitor," *MIT Energy Laboratory Technical Report, and Electric Power Research Institute Technical Report*, 1985.

[5] M. Kaselimi, E. Protopapadakis, A. Voulodimos, N. Doulamis, and A. Doulamis, "Towards trustworthy energy disaggregation: A review of challenges, methods, and perspectives for non-intrusive load monitoring," *Sensors*, vol. 22, no. 15, p. 5872, 2022.

[6] L. Pereira and N. J. Nunes, "Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—a review," *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 8, no. 6, p. e1265, 2018.

[7] L. Pereira and N. Nunes, "A comparison of performance metrics for event classification in non-intrusive load monitoring," in *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2017, pp. 159–164.

[8] C. Nalmpantis and D. Vrakas, "Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 217–243, 2019.

[9] C. Klemenjak, S. Makonin, and W. Elmenreich, "Towards comparability in non-intrusive load monitoring: On data and performance evaluation," in *2020 IEEE power & energy society innovative smart grid technologies conference (ISGT)*. IEEE, 2020, pp. 1–5.

[10] L. Pereira and N. Nunes, "An empirical exploration of performance metrics for event detection algorithms in non-intrusive load monitoring," *Sustainable Cities and Society*, vol. 62, p. 102399, 2020.

[11] D. Garcia-Perez, D. Pérez-López, I. Diaz-Blanco, A. González-Muñiz, M. Domínguez-González, and A. A. C. Vega, "Fully-convolutional denoising auto-encoders for nilm in large non-residential buildings," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2722–2731, 2020.

[12] E. T. Mayhorn, G. P. Sullivan, J. M. Petersen, R. S. Butner, and E. M. Johnson, "Load disaggregation technologies: real world and laboratory performance," *Pacific Northwest National Laboratory (PNNL), Richland, WA (US), Tech. Rep. PNNL-SA-116560*, 2016.

[13] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of united kingdom households from a two-year longitudinal study," *Scientific data*, vol. 4, no. 1, pp. 1–12, 2017.

[14] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, "Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey," *Sensors*, vol. 12, no. 12, pp. 16 838–16 866, 2012.

[15] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[16] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," in *Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA*, vol. 25, no. Citeseer. Citeseer, 2011, pp. 59–62.

[17] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han, "Unsupervised disaggregation of low frequency power measurements," in *Proceedings of the 2011 SIAM international conference on data mining*. SIAM, 2011, pp. 747–758.

[18] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "Nilmtk: An open source toolkit for non-intrusive load monitoring," in *Proceedings of the 5th international conference on Future energy systems*, 2014, pp. 265–276.

[19] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson, "Towards reproducible state-of-the-art energy disaggregation," in *Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation*, 2019, pp. 193–202.

[20] S. Makonin and F. Popowich, "Nonintrusive load monitoring (nilm) performance evaluation," *Energy Efficiency*, vol. 8, no. 4, pp. 809–814, 2015.

[21] N. Zaeri, A. Ashouri, H. B. Gunay, and T. Abuimara, "Disaggregation of electricity and heating consumption in commercial buildings with building automation system data," *Energy and Buildings*, vol. 258, p. 111791, 2022.

[22] S. Sahrane and M. Haddadi, "Near real-time low frequency load disaggregation," *ENP Engineering Science Journal*, vol. 1, no. 2, pp. 50–54, 2021.

[23] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. Griffin, 1973.

[24] E. F. Glynn, "Correlation "distances" and hierarchical clustering," 2005.

[25] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[26] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985.

[27] https://www.lg.com/africa/blog/difference-between-refrigerator-and-freezer, accessed: 2024-10-20.