CAU Kiel University Christian-Albrechts-Universität zu Kiel

IEEE PES Power & Energy Society

IEEE Advancing Technology for Humanity

PowerTech KIEL 2025

June 29th - July 3rd, 2025

Empirical Exploration of Energy Estimation Metrics in Load Disaggregation

João Góis, Lucas Pereira

ITI/LARSyS, Instituto Superior Técnico, University of Lisbon

MOTIVATION

Improving energy consumption feedback in buildings helps mitigate climate change

Non-Intrusive Load Monitoring (NILM) estimates appliance-level energy use. However, NILM algorithms' performance depends on the chosen evaluation metrics. This aspect is still underexplored, especially across appliance types [1, 2, 3].

GOAL

Empirically analyze how relationships between Energy Estimation (EE) metrics vary across different appliance types, offering improved guidance on metric selection in NILM.

EMPIRICAL APPROACH

Dataset

REFIT (2 years of household energy data, 8 appliances).

Algorithms

4 NILM baselines (Edge Detection, Combinatorial Optimization, Mean and Exact FHMM).

Metrics

15 EE metrics, ranging from traditional metrics from other fields to those specific to NILM:

- Average Error (AE)
- Absolute Error (ABSE)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Coefficient of Variation of RMSE (CVRMSD)
- Energy Error (EE)
- Energy Accuracy (EA)
- Match Rate (MR)
- Fraction of Energy Explained (FEE)
- R-squared (R^2)
- Percentage of Standard Deviation Explained (PSDE)
- Standard Deviation of the Error (SDE)
- Standard Error of the Mean (SEM)
- Error in Total Assigned Energy (ETEA)
- Deviation (Dev)

Analysis

- Spearman correlation between metrics
- Hierarchical clustering of metrics for each appliance
- Rand Index to compare metric groupings across appliances

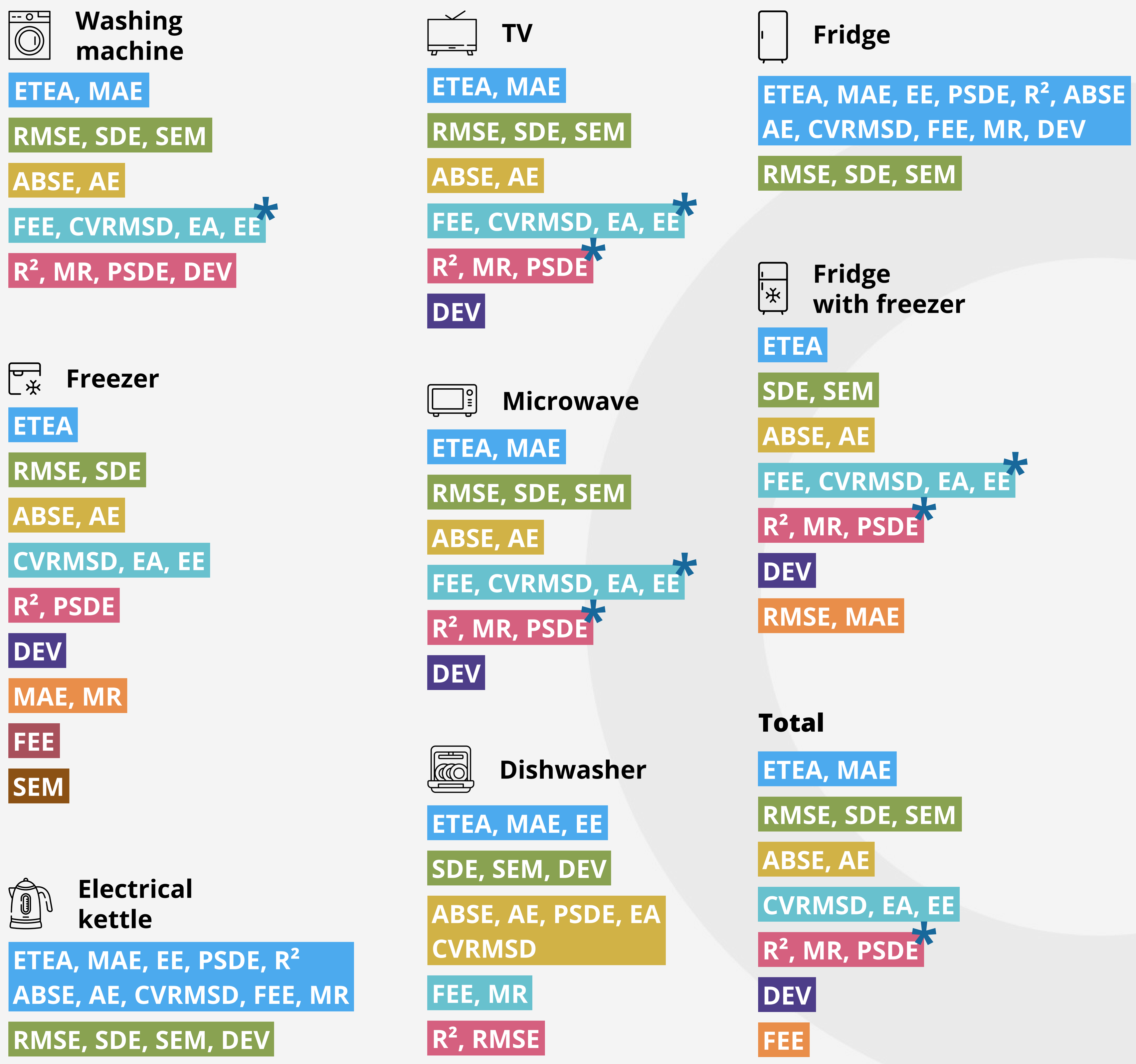

Figure 1: Cluster partition of the metrics across appliances and all-appliances set.

KEY FINDINGS

Metrics group differently across appliance types

Multistate appliances (washing machine, microwave, fridge freezer, dishwasher, and freezer) have similar clusters. The kettle (on/off appliance) shows a different metric clustering.

Consistent metric groups found

- FEE, CVRMSD, EA, EE*
- R^2, MR, PSDE*

EA and MR are robust, normalized metrics for comparison (in line with [1]).

CONCLUSION

MR and EA are promising metrics for consistent NILM evaluation. This supports more meaningful performance comparisons and improve energy feedback in real-world scenarios.

ACKNOWLEDGEMENTS

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

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

[3] Pereira, L., & Nunes, N. (2020). An empirical exploration of performance metrics for event detection algorithms in non-intrusive load monitoring. Sustainable Cities and Society, 62, 102399.

joao.gois@iti.larsys.pt

Interactive Technologies Institute

Técnico Lisboa