Identifying Impactful Devices on Disaggregation Performance

Sean Barker, Anna Leitner, and Andy Stoneman
Bowdoin College
[sbarker,aleitner,astonema]@bowdoin.edu

ABSTRACT

One of the key barriers to widespread deployment of disaggregation algorithms is the difficulty that these algorithms have in real-world environments containing many devices. While a greater number of devices inevitably results in "noisier" aggregate consumption, different devices will have varying impacts on the overall difficulty of disaggregation. To investigate the extent of this effect, we conduct an empirical study in which we disaggregate fixed appliance loads from real-world data while systematically varying the set and complexity of other loads present. Our study highlights the outsized impact of certain device types and suggests paths towards scaling disaggregation algorithms to more complex environments.

CCS CONCEPTS

• **General and reference** → **Performance**; **Evaluation**; **Empirical studies**; • **Hardware** → **Energy metering**.

KEYWORDS

Energy disaggregation, NILM scalability, background devices, noise

ACM Reference Format:
Sean Barker, Anna Leitner, and Andy Stoneman. 2022. Identifying Impactful Devices on Disaggregation Performance. In *6th International Workshop on Non-Intrusive Load Monitoring (NILM '22), November 9–10, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3563357.3566150

1 INTRODUCTION

Energy disaggregation (also known as non-intrusive load monitoring or NILM) refers to decomposing aggregate energy consumption collected at the building level into all of its component loads [4]. Ideally, the sum of the resulting disaggregated loads equals the aggregate consumption, providing a complete breakdown of energy usage. However, owing to the large number of electrical devices present in most buildings and the difficulty of accurately modeling all of them, it is often impractical in real environments to provide a full breakdown. Disaggregation often focuses instead on specific devices of interest, such as large appliances or power-intensive devices. In doing so, disaggregation can provide useful insights without having to handle the large number of other devices present. In this work, we refer to non-disaggregated devices as *background devices*, which have also been termed 'standby loads' or simply 'noise' – the latter referring to consumption from unknown sources [3, 12]. We similarly term the devices to be disaggregated *foreground devices.* While simple disaggregation scenarios may involve only foreground devices, most real-world scenarios will also involve some number of background devices.

The presence of background devices will clearly have some impact on disaggregation performance. However, the extent of this impact is unlikely to be linear; for example, a background device that is highly active is likely to have more impact than a background device that is rarely used. Thus, disaggregation difficulty may stem from specific, highly impactful devices rather than from a large number of total devices. Investigating the impact of specific devices can inform how to scale NILM to noisy, real-world environments.

In this paper, we explicitly consider the impact of background devices on disaggregation performance by conducting an empirical study on building environments of systematically varied complexity. In doing so, we aim to increase our understanding of practical NILM scalability. Our study explores the impact of background devices on widely-used disaggregation algorithms and points to the outsized impact of some of these devices.

2 METHODOLOGY

The primary goal of our study is to identify and characterize the devices most impactful on disaggregation – particularly devices that may themselves not be part of the disaggregation (i.e., background devices). Formally, consider a building containing the set of devices $D = \{d_1, d_2, ..., d_n\}$. The set D is decomposed into the set F of foreground devices to be disaggregated and the remaining set B of background devices – that is, $D = B \cup F$. Devices in B are part of the aggregate consumption but are not themselves disaggregated and may not even be known to the disaggregation algorithm. While the classic definition of NILM consists exclusively of foreground devices ($D = F$), in practice, studies often focus on a small set of foreground devices (typically larger appliances) [5]. To assess the impact of background devices, we fix F while varying B. As B grows, the aggregate data becomes noisier and degrades performance. While disaggregation studies such as [10] perform a similar operation by injecting artificial noise into the aggregate trace, we instead modify disaggregation complexity via the background device set.

2.1 Ranking Device Impact

A simple approach to investigating the relative impact of different devices is to progressively add background devices to B, re-running disaggregation on the same foreground devices with each addition to measure the effect. However, such an approach is of limited utility, as the impact of adding a given device is highly dependent on the devices already added (or not yet added). We instead formulate the problem by considering a complete device set D from which we wish to identify the most impactful k devices. We can find the optimal set of k devices by considering every possible set of k devices

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

NILM '22, November 9–10, 2022, Boston, MA, USA

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9890-9/22/11…$15.00
https://doi.org/10.1145/3563357.3566150

and re-testing with each set held back from D. Unfortunately, this optimal approach is computationally intractable for larger numbers of devices – e.g., for $|D| = 50$ and $k = 5$, there are over a million possible combinations of the 45 background devices to consider.

We instead propose a greedy algorithm to approximate the top devices. As outlined in Algorithm 1, for a given starting configuration of foreground set F and background set B, we run $|B|$ disaggregation trials in which we remove a single device from B (i.e., leaving $|B| - 1$ background devices). In each trial, we disaggregate the devices in F and then score the results relative to the baseline where all devices in B are included. The highest-scoring trial determines the single next device to remove from B (i.e., the most impactful device) during subsequent rounds. With this approach, we can estimate the top k devices using only $O(|B| \times k)$ disaggregation trials.

Algorithm 1 Disaggregation Impact Ranking Algorithm

```
1:  procedure FindTopK($F, B, k$)
2:      $topK \leftarrow []$
3:      for $i$ from 1 to $k$ do                          ▷ for each of $k$ rounds
4:          $base \leftarrow disaggregate(F, B)$           ▷ current round baseline
5:          $scores \leftarrow []$
6:          for each $device$ in $B$ do
7:              $B.remove(device)$
8:              $result \leftarrow disaggregate(F, B)$     ▷ re-test on $F$
9:              $score \leftarrow score(result, base, F)$  ▷ scoring function
10:             $scores.add(score)$
11:             $B.add(device)$                            ▷ replace device
12:         $scores.sort()$
13:         $topK.add(scores[0])$                          ▷ top-scoring device
14:         $B.remove(top)$                                ▷ remove from future rounds
15:     return $topK$
```

2.2 Scoring Results

Algorithm 1 does not specify how the scoring function works (line 9) – i.e., how do we compute the improvement in disaggregating the foreground device set? This problem is nontrivial owing to the lack of standardization and multiple metrics used in NILM performance measurement [6–9]. Here, we optimize a single (configurable) metric, as follows: for any standard disaggregation metric (MAE, RMSE, F1 score, etc.), we sum the absolute improvement versus the prior baseline across all devices in F. This sum excludes any devices that see degraded performance, as these cases can be safely ascribed to algorithmic randomness or noise. Although our present scoring function uses only a single metric at once, more sophisticated approaches could consider multiple metrics at once, particularly in complimentary ways (such as proposed in [7]).

3 ALGORITHMS, DATA, AND METRICS

We conduct our study using the latest version of the NILMTK disaggregation toolkit [1, 2]. Since our aim is not to focus on any specific disaggregation algorithm, we run our experiments using a variety of classic and cutting-edge NILM algorithms implemented in NILMTK: Edge Detection (Edge), Combinatorial Optimization (CO), Exact Factorial Hidden Markov Model (FHMM), Denoising Autoencoder (DAE), Sequence-to-Sequence (S2S), Sequence-to-Point (S2P), and Recurrent Neural Network (RNN). Neural networks are trained for 50 epochs with a batch size of 128. All other algorithm-specific parameters are left at their default values within NILMTK.

Our experiments are conducted using real-world device data from Dataport [11] with a sampling rate of 10. Since this dataset is circuit-level rather than device-level (as with most larger datasets), there is the possibility of multi-device circuits, but most typical appliances and larger devices exist on dedicated circuits. As such, we treat circuits and devices equivalently in this study.

We are not concerned with the specific building from which each device originates, but simply treat the dataset as a repository of device traces from which to construct test buildings. We thus discard the Dataport aggregates in favor of synthetic aggregates constructed by summing all foreground and background devices under consideration. Our total device set consists of 35 circuits selected to encompass all common device types in the dataset (typical appliances, lighting, HVAC, etc.) while ensuring that each device exhibits some activity over the month-long training and testing periods (two months of data in total). Each of our experiments is run on a synthetic house containing a subset of the 35 circuits.

We focus on two complimentary evaluation metrics, following the example of [10] and others. First, we consider mean absolute error (MAE), which measures the predictive error between the disaggregated consumption over time and the real consumption. Second, we consider F1 score, which labels each time period using a binary active/inactive classification and measures the accuracy of the disaggregated labeling versus ground truth. Both are standard metrics widely used in disaggregation work [7].

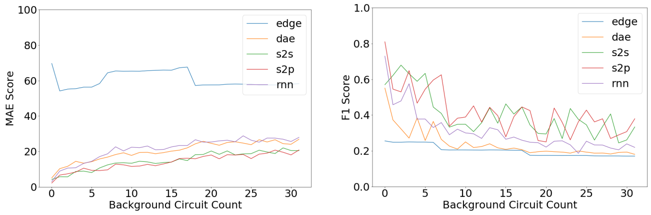
4 RESULTS

As detailed in Section 2, each of our experiments is configured as a particular set of foreground devices to disaggregate in the presence of zero or more background devices. We focus on disaggregation performance of four typical appliances: a refrigerator, dishwasher, microwave, and washing machine. These device types are popular in disaggregation studies and are varied in their usage patterns (e.g., a refrigerator is frequent and regular, while a microwave is infrequent and less predictable). These four devices comprise the foreground devices in all experiments while we vary the background device set. We consider two baseline configurations: no background devices (the "easiest" case), which we term **BG-0**, and all 31 background devices (the "hardest" case), which we term **BG-31**.

In all experiments, the disaggregation performance of FHMM and CO was poor relative to all other algorithms. Hence, for brevity and clarity, we only report results from the other five disaggregation algorithms (Edge, DAE, S2S, S2P, and RNN).

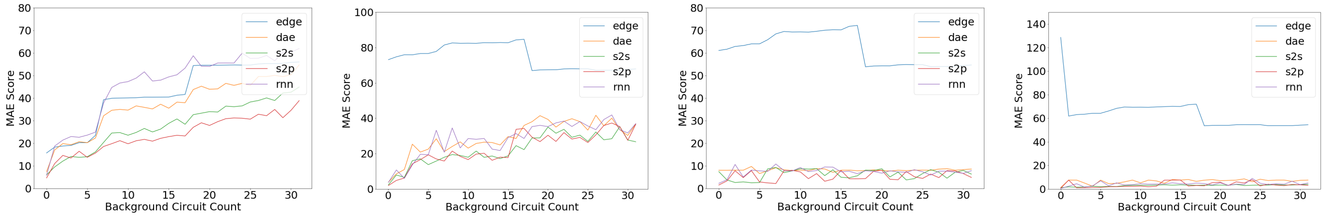
4.1 Scaling Disaggregation Complexity

We first disaggregate the **BG-0** configuration and then progressively add background devices until arriving at **BG-31**, re-running disaggregation with each added device. The order of added devices is chosen arbitrarily. Figure 1 shows the averaged MAE and F1 scores across all foreground devices for each test run. We see that all algorithms experience a continuous but relatively smooth degradation as background activity increases, though the magnitude varies



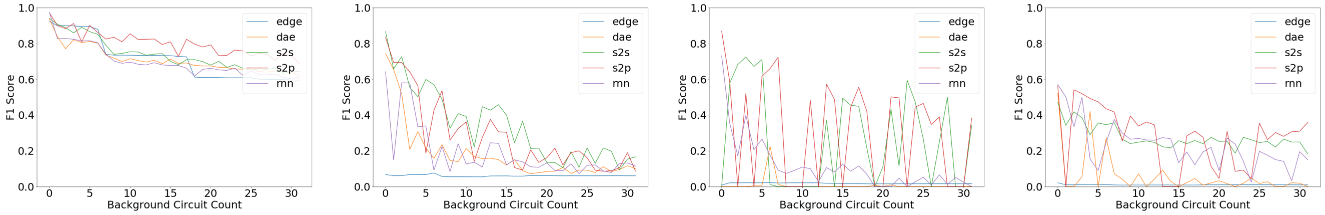
(a) MAE (b) F1 Score

Figure 1: Average MAE and F1 score for all foreground devices as background activity increases.



(a) Refrigerator (b) Dishwasher (c) Washing Machine (d) Microwave

Figure 2: MAE for individual foreground devices as background activity increases.



(a) Refrigerator (b) Dishwasher (c) Washing Machine (d) Microwave

Figure 3: F1 scores for individual foreground devices as background activity increases.

substantially by device – in particular, the refrigerator and dishwasher appear to be most sensitive. The simple Edge algorithm is reliably the lowest performer, particularly in noisier environments, while the other approaches (all based on neural nets) are generally competitive. However, the best-performing algorithm (S2P) still experiences a 59% drop in average F1 score and 8X increase in average MAE between BG-0 and BG-31.

Figures 2 and 3 depict MAE and F1 scores for each individual foreground device across each test run. We see that the refrigerator is relatively easy to disaggregate (even using simple edge detection), but experiences the most notable degradation with increasing complexity. The dishwasher also exhibits fairly smooth degradation. In contrast, the washing machine and microwave are harder to disaggregate in general, but display less pronounced trends across runs (i.e., the addition of background devices less consistently degrades performance). These results suggest that while highly-accurate disaggregation is difficult in noisy or complex environments, maintaining a lower but more stable level of performance may be possible even in noisy environments.

4.2 Most Impactful Devices

We next employ the ranking algorithm described in Section 2.2 to estimate the five most impactful background devices from the entire set. We consider the nine different ranking configurations shown in Table 1 to choose the top five devices. Each configuration is described by (1) which algorithm (S2P or Edge) is used to disaggregate in each round, (2) the number of rounds used – in the case of a single round, the top five devices are chosen immediately based on the results of the first round, and (3) which metric (MAE or F1) is used to produce the ranking in each round. We also consider a manual configuration (denoted TYPE), in which we manually select devices matching similar devices type to the foreground devices.

We see from the resulting device sets that there are significant variations across configurations, though several devices are highly ranked by most or all configurations (e.g., `AirwindowUnit-5746` and `Freezer-142`). The device type does not appear to be strongly predictive of a high ranking, as some background devices of the same type as a foreground device (e.g., `Refrigerator-3700` are not ranked except by TYPE. We also see that picking all 5 devices from only one round can produce a notably different set of devices – for example, SP-5F and SP-1F share only two of five top devices.

To evaluate the quality of each configuration and the impact of the chosen devices, we remove the top five ranked devices from the full set of 31 background devices and then disaggregate the foreground devices with the remaining 26 background devices. Owing to the globally superior disaggregation performance of S2P versus Edge, S2P is used to perform the disaggregation for each configuration regardless of whether S2P or Edge was used to choose

Name	Rank Alg	Rounds	Metric	Top Five Devices (ranked)
SP-5M	S2P	5	MAE	LightsPlug-2096, Air-2361, AirwindowUnit-5746, Freezer-142, Heater-2318
SP-1M	S2P	1	MAE	LightsPlug-2096, AirwindowUnit-5746, Air-2361, Car-1222, WellPump-1222
SP-5F	S2P	5	F1 Score	Microwave-3700, Air-2361, Heater-2318, Bedroom-5746, Furnace-3488
SP-1F	S2P	1	F1 Score	Microwave-3700, Car-1222, AirwindowUnit-5746, Housefan-5058, Heater-2318
ED-5M	Edge	5	MAE	Freezer-142, AirwindowUnit-5746, Dryer-1222, Air-2361, WellPump-1222
ED-1M	Edge	1	MAE	Freezer-142, AirwindowUnit-5746, Dryer-1222, Air-2361, Garage-4373
ED-5F	Edge	5	F1 Score	AirwindowUnit-5746, Freezer-142, Air-2361, LightsPlug-2096, Furnace-3488
ED-1F	Edge	1	F1 Score	AirwindowUnit-5746, Freezer-142, LightsPlugs-2096, Air-2361, Microwave-3700
TYPE	Manual	N/A	Device Type	AirwindowUnit-5746, Freezer-142, Microwave-3700, Refrigerator-3700, Range-558

Table 1: Ranking configurations and corresponding top five devices (in ranked order for each configuration).

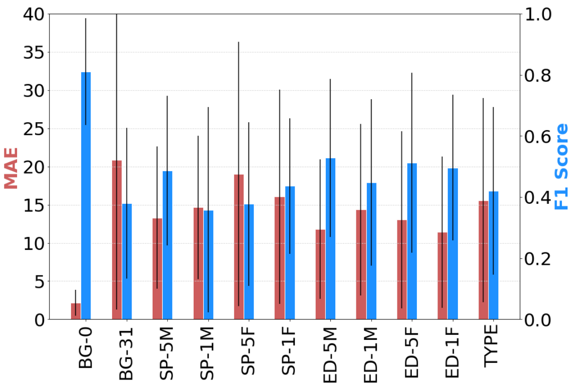
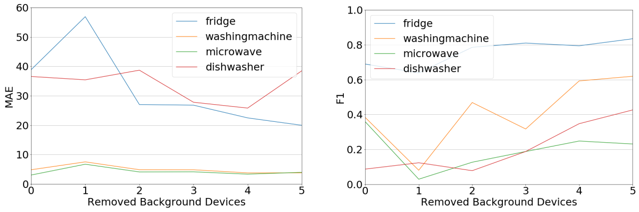


Figure 4: Average MAE and F1 scores across all foreground devices for baselines and each ranking configuration from Table 1 when disaggregated using S2P.



(a) MAE

(b) F1 Score

Figure 5: Individual MAE and F1 scores for S2P disaggregation using ED-5M ranking configuration.

the top devices for that configuration. Figure 4 depicts the average MAE and F1 scores across all foreground devices for each of the nine configurations plus the two baselines BG-0 and BG-31.

Interestingly, we see that the most impactful (i.e., "best") ranking is produced by the ED-5M configuration, and more generally, that the rankings produced using Edge are significantly superior to those produced using S2P (even when disaggregation is ultimately performed using S2P rather than Edge, as is the case in all results shown in Figure 4). We suspect that this is due to the relative simplicity of Edge; only significant and reliable improvements are likely to be captured at all by Edge rankings, and hence, the devices chosen by these rankings are strongly impactful even when a more sophisticated disaggregation algorithm is ultimately used. In all cases other than SP-5F, all of the 5-round configurations are superior to their otherwise-identical 1-round counterparts.

We also consider the impact of the most significant devices on individual foreground devices. Figure 5 shows the individual MAE and F1 scores for S2P disaggregation as the top devices chosen by ED-5M are removed one at a time (in order of significance). We see significant improvements in F1, MAE, or both for all devices except for the microwave, which sees minimal improvement from a poor starting baseline. This result reflects that while some devices may be impacted primarily by a few specific other devices and thus easily improved, other devices may be more inherently difficult to recognize irrespective of specific background activity.

An important question is the extent to which removing the "top" devices bridges the performance distance between the BG-31 and BG-0 baselines. Numerically, removing the top five devices chosen by ED-5M (representing only 14% of the total devices) eliminates 40% of the overall F1 delta and 48% of the overall MAE delta between BG-31 and BG-0. These results reflect the outsized influence that these few devices have on the complexity of the aggregate data. Extending the ED-5M configuration out to 10 rounds (i.e., ED-10M) and removing the resulting top 10 devices eliminated 65% of the F1 delta and 66% of the MAE delta between BG-31 and BG-0.

Qualitatively, most of the top-ranked devices (across all configurations, as listed in Table 1) are characterized by frequent activity (e.g., cyclic devices such as freezers and air conditioners are heavily represented), and, to a lesser extent, high power consumption. As our experiments demonstrate, these devices are significantly more impactful on disaggregation than average, and removing them from a device set to be disaggregated (such as by installing a small number of dedicated plug-level meters on such devices) may have an outsized impact on the feasibility of disaggregating other devices.

5 CONCLUSION

Performing accurate disaggregation in buildings containing large numbers of devices remains a difficult problem. This study explores the extent to which this difficulty may stem from specific, highly impactful devices, as opposed to simply the total number of devices present. Our results show that a small number of cyclic, frequently operating devices may have an outsized impact on disaggregation performance, and thus are especially important to consider when scaling NILM to larger homes and buildings. As future work building on this study, we intend to characterize highly impactful devices in more depth and explore practical, minimally-invasive techniques to identify such devices in novel operating environments. This work represents a step towards effectively deploying NILM algorithms in real-world buildings containing many and varied devices.

REFERENCES

[1] Nipun Batra, Jack Kelly, Oliver Parson, Haimonti Dutta, William Knottenbelt, Alex Rogers, Amarjeet Singh, and Mani Srivastava. 2014. NILMTK: An Open Source Toolkit for Non-Intrusive Load Monitoring. In *Proceedings of the 5th International Conference on Future Energy Systems* (Cambridge, United Kingdom) *(e-Energy ’14)*. Association for Computing Machinery, New York, NY, USA, 265–276. https://doi.org/10.1145/2602044.2602051

[2] Nipun Batra, Rithwik Kukunuri, Ayush Pandey, Raktim Malakar, Rajat Kumar, Odysseas Krystalakos, Mingjun Zhong, Paulo Meira, and Oliver Parson. 2019. 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* (New York, NY, USA) *(BuildSys ?19)*. Association for Computing Machinery, New York, NY, USA, 193?202. https://doi.org/10.1145/3360322.3360844

[3] Dominik Egarter, Manfred Pöchacker, and Wilfried Elmenreich. 2015. Complexity of Power Draws for Load Disaggregation. *CoRR* abs/1501.02954 (2015). arXiv:1501.02954 http://arxiv.org/abs/1501.02954

[4] G.W. Hart. 1992. Nonintrusive Appliance Load Monitoring. *Proc. IEEE* 80, 12 (Dec. 1992), 1870–1891. https://doi.org/10.1109/5.192069

[5] Jack Kelly and William Knottenbelt. 2015. Neural NILM: Deep Neural Networks Applied to Energy Disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments* (Seoul, South Korea) *(BuildSys ’15)*. Association for Computing Machinery, New York, NY, USA, 55?64. https://doi.org/10.1145/2821650.2821672

[6] Christoph Klemenjak, Stephen Makonin, and Wilfried Elmenreich. 2020. 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.

[7] Stephen Makonin and Fred Popowich. 2015. Nonintrusive load monitoring (NILM) performance evaluation. *Energy Efficiency* 8, 4 (2015), 809–814. https://doi.org/10.1007/s12053-014-9306-2

[8] L. Pereira and N. Nunes. 2017. A comparison of performance metrics for event classification in Non-Intrusive Load Monitoring. In *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. 159–164.

[9] L. Pereira and N. Nunes. 2018. An Experimental Comparison of Performance Metrics for Event Detection Algorithms in NILM. In *4th International Workshop on Non-Intrusive Load Monitoring (NILM 2018)*.

[10] Andreas Reinhardt and Christoph Klemenjak. 2020. How Does Load Disaggregation Performance Depend on Data Characteristics? Insights from a Benchmarking Study. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems* (Virtual Event, Australia) *(e-Energy ?20)*. Association for Computing Machinery, New York, NY, USA, 167?177. https://doi.org/10.1145/3396851.3397691

[11] Pecan Street. (2019). Dataport. https://dataport.pecanstreet.org/. Accessed January 2020.

[12] B. Zhao, K. He, L. Stankovic, and V. Stankovic. 2018. Improving Event-Based Non-Intrusive Load Monitoring Using Graph Signal Processing. *IEEE Access* 6 (2018), 53944–53959.