Energy Informatics

SHORT COMMUNICATION Open Access

Investigating the performance gap between testing on real and denoised aggregates in non-intrusive load monitoring

Christoph Klemenjak[1]*, Stephen Makonin[2] and Wilfried Elmenreich[1]

*Correspondence:
klemenjak@ieee.org
[1]Institute of Networked and Embedded Systems, University of Klagenfurt, Klagenfurt, Austria
Full list of author information is available at the end of the article

Abstract

Prudent and meaningful performance evaluation of algorithms is essential for the progression of any research field. In the field of Non-Intrusive Load Monitoring (NILM), performance evaluation can be conducted on real-world aggregate signals, provided by smart energy meters or artificial superpositions of individual load signals (i.e., denoised aggregates). It has long been suspected that testing on these denoised aggregates provides better evaluation results mainly due to the fact that the signal is less complex. Complexity in real-world aggregate signals increases with the number of unknown/untracked loads. Although this is a known performance reporting problem, an investigation into the actual performance gap between real and denoised testing is still pending. In this paper, we examine the performance gap between testing on real-world and denoised aggregates with the aim of bringing clarity into this matter. Starting with an assessment of noise levels in datasets, we find significant differences in test cases. We give broad insights into our evaluation setup comprising three load disaggregation algorithms, two of them relying on neural network architectures. The results presented in this paper, based on studies covering three scenarios with ascending noise levels, show a strong tendency towards load disaggregation algorithms providing significantly better performance on denoised aggregate signals. A closer look at the outcome of our studies reveals that all appliance types could be subject to this phenomenon. We conclude the paper by discussing aspects that could be causing these considerable gaps between real and denoised testing in NILM.

Keywords: Load disaggregation, Non-intrusive load monitoring, Denoised testing, Performance evaluation, Energy datasets

Introduction

Effective energy management in smart grids requires a fair amount of monitoring and controlling of electrical load to achieve optimal energy utilization and, ultimately, reduce energy consumption (Gopinath et al. 2020). With regard to individual buildings, load monitoring can be implemented in an intrusive or non-intrusive fashion. The latter is often referred to as Non-Intrusive Load Monitoring (NILM) or load disaggregation. NILM, dating back to the seminal work presented in Hart (1992), comprises a set of

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

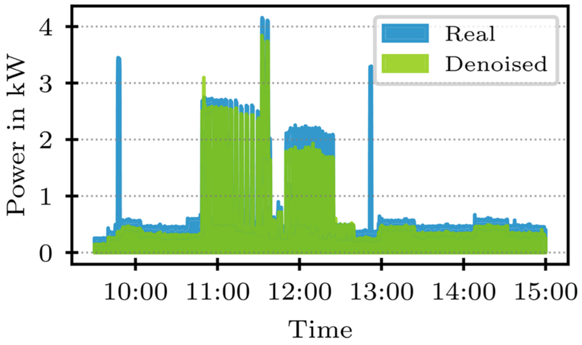

Fig. 1 Real and denoised aggregate in the case of UK-DALE house 5

techniques to identify active electrical appliance signals from the aggregate load signal reported by a smart meter (Salem et al. 2020).

Performance evaluation of NILM algorithms can be carried out in a noised or denoised manner, where the difference lies in the aggregate signal considered as input. Whereas noised scenarios employ signals (i.e. time series) obtained from smart meters, denoised testing scenarios consider superpositions of individual appliance signals (i.e., denoised aggregates). Figures 1 and 2 illustrate such real and denoised signals for two households found in NILM datasets. Depending on how many appliance signals are considered when deriving a denoised aggregate, there can be test scenarios in which the denoised aggregate differs considerably from its real-world counterpart, as shown in Fig. 2.

While a large proportion of contributions proposed for NILM is being evaluated following noised testing scenarios, exceptions to this unwritten rule can be observed (Wittmann et al. 2018). The problem with this matter lies in the complexity of the test setup, as denoised aggregates are suspected to pose simpler disaggregation problems (Makonin and Popowich 2015). Consequently, our hypothesis claims that the same disaggregation algorithm applied to the denoised signal version of a real-world aggregate signal results in considerably better performance, thus communicating a distorted picture of the capabilities of the presented algorithm.

This paper presents a study focusing on the difference between denoised and real-world signal testing scenarios in the context of performance evaluation in NILM. We consider data of 15 appliances extracted from three datasets. Each dataset reports an aggregate

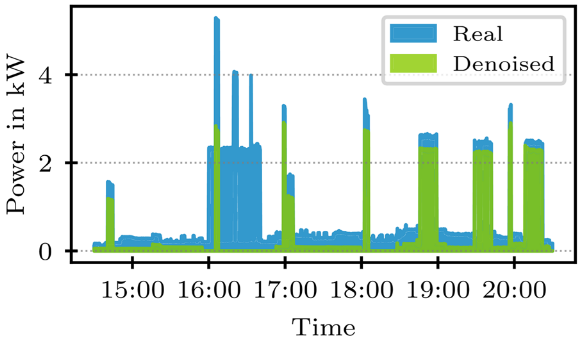

Fig. 2 Real and denoised aggregate in the case of REFIT house 2

signal with additional residual noise. For testing, we select households with different levels of residual noise. We incorporate one basic and two load disaggregation approaches based on neural networks to obtain a broad understanding of whether or not noise levels of aggregate power signals impact energy estimation performance. Finally, we discuss how the disaggregation performance is affected by signal noise levels with regard to different appliance types.

Related work

Despite the possibly far-reaching implications of this aspect for NILM, relatively little is understood about the actual performance gap between real and denoised testing. In Makonin and Popowich (2015), the hypothesis of denoised testing resulting in better performance was expressed first. Further, the authors introduce a measure to assess the noise level of aggregate signals. This measure has found application in a limited number of studies, in which the noise level was reported alongside the performance of load disaggregation algorithms on real-world aggregates (Makonin et al. 2015; Zhao et al. 2018). However, no comparison to the denoised testing case has been conducted. In Klemenjak et al. (2020), the noise levels of several NILM datasets were determined. The authors report basic parameters of several NILM datasets and find that noise levels in real aggregate signals vary significantly among observed datasets.

Few attempts have been made to evaluate NILM algorithms on both, real and denoised aggregates, such as presented for the AFAMAP approach in Bonfigli et al. (2017). In subsequent work (Bonfigli et al. 2018), an improved version of denoising autoencoders for NILM has been proposed by means of comparison studies to the state of the art at that time. Although the authors have not investigated the performance gap between real and denoised, a tendency can be derived for this particular case in both contributions, confirming the motivation for the studies presented in this paper.

Assessing signal noise levels

NILM has been approached in various ways that can be categorized into event detection and energy estimation approaches (Pereira and Nunes 2018). In the following, we focus on the energy estimation viewpoint as the precursor of the event detection stage in the disaggregation process. We define NILM as the problem of generating estimates $\left[\hat{x}_t^{(1)}, \ldots, \hat{x}_t^{(M)}\right]$ of the actual power consumption $\left[x_t^{(1)}, \ldots, x_t^{(M)}\right]$ of M electrical appliances at time t given only the aggregated power consumption y_t, where the aggregate power signal y_t consists of

$$y_t = \sum_{i=1}^{M} x_t^{(i)} + \eta_t \tag{1}$$

that is M appliance-level signals $x_t^{(i)}$ and a residual term η_t. The residual term comprises (measurement) noise, unmetered electrical load, and unexpected or unaccounted anomalies (Makonin and Popowich 2015). To quantify the share of the residual term in an aggregate signal, the noise-aggregate ratio NAR, defined as:

$$\text{NAR} = \frac{\sum_{t=1}^{T} \eta_t}{\sum_{t=1}^{T} y_t} = \frac{\sum_{t=1}^{T} |y_t - \sum_{i=1}^{M} x_t^{(i)}|}{\sum_{t=1}^{T} y_t} \tag{2}$$

was introduced in Makonin and Popowich (2015). This ratio can be computed for any type of power signal, provided that readings of the aggregate and individual appliances are available. A NAR of 0.15 indicates that 15% of the observed power signal can be attributed to the residual term. Hence, the ratio indicates to what degree information on the aggregate's components is available.

To get an impression of NAR levels to be expected in real-world settings, we compute this ratio for household measurements contained in the energy datasets AMPds2 (Makonin et al. 2016), COMBED (Batra et al. 2014a), ECO (Beckel et al. 2014), iAWE (Batra et al. 2013), REFIT (Murray et al. 2017), and UK-DALE (Kelly and Knottenbelt 2015a). As intended by the authors of (Makonin and Popowich 2015), we consider all submeter signals recorded during the measurement campaign to compute the NAR. These datasets were selected because of their compatibility to NILMTK, a toolkit that enables reproducible NILM experiments (Batra et al. 2014b; Batra et al. 2019). We excluded from consideration the dataset BLUED (Anderson et al. 2012) due to the lack of sub-metered power data, Tracebase (Reinhardt et al. 2012) and GREEND (Monacchi et al. 2014) due to the lack of household aggregate power data. We summarize the derived values in Table 1 in conjunction with further stats on the measurement campaign such as duration or number of submeters.

Generally speaking, measurement campaigns strive to record the energy consumption and other parameters of interest in households or industrial facilities over a certain time period. Though sharing this common aim, considerable differences can be observed in the way past campaigns have been conducted. As Table 1 shows, durations range from a couple of days to several years of data, which impacts the amount of appliance activations and events found in the final dataset. Further, we identify considerable variations with regard to AC power types as well as the number of submeters installed during campaigns.

Table 1 Noise levels in NILM datasets

Dataset	**House**	**Duration**	**Meters**	**Power Types**		**NAR**
		[days]				**[%]**
AMPds2	1	730	20	P, Q, S	P, Q, S	17.8
COMBED	1	28	13	P	P	34.4
ECO	1	236	7	P, Q	P	67.0
ECO	2	245	12	P, Q	P	5.9
ECO	3	57	7	P, Q	P	97.0
ECO	4	211	8	P, Q	P	70.5
ECO	5	219	8	P, Q	P	84.7
ECO	6	124	7	P, Q	P	66.0
iAWE	1	60	10	P, Q, S	P, Q, S	50.0
REFIT	1	639	9	P	P	64.5
REFIT	2	617	9	P	P	65.1
REFIT	3	614	9	P	P	55.5
REFIT	4	634	9	P	P	52.5
REFIT	5	648	9	P	P	52.3
UK-DALE	1	658	52	P, S	P, S	33.3
UK-DALE	2	110	18	P, S	P	41.2
UK-DALE	3	35	4	S	P	-
UK-DALE	4	114	5	S	P	-
UK-DALE	5	107	24	P, S	P	27.5

Table 2 Details on intervals and dataset splitting

Dataset	House	Interval	Train [days]	Test [days]
ECO	2	2012-06-01 to 2013-01-31	207	37
REFIT	2	2014-03-01 to 2014-12-01	234	41
UK-DALE	5	2014-07-25 to 2014-10-15	70	12

It should be pointed out that there seems to be a lack of consistency in the sense that not only measurement setups differ between two datasets but also within some of the campaigns considered by our comparison (e.g., UK-DALE).

As concerns the noise aggregate ratio (NAR), we observe considerable variations across datasets and households. Interestingly, the NAR ranges between a few percent, as it is the case for household 2 in the ECO dataset, and excessive 84.7% in household 5 of same dataset. Further, there are indications that the number of submeters used in the course of dataset collection can but do not necessarily have an impact on the noise level of the household's aggregate signal since it is decisive what kind of appliances are left out during a measurement campaign (low-power appliances vs. big consumers). As concerns house 1 to house 5 in REFIT, we consistently observe moderate to high noise levels, which may be the result of the low number of submeters incorporated in the measurement campaign. On the other hand, it should be noted that the measurement campaign conducted to obtain REFIT shows remarkable consistency in the sense that the exact same number of submeters has been applied to every single household in the study and, more importantly, the same AC power type has been considered at aggregate and appliance level at every site. In contrast to that, Table 1 reveals that in the case of house 3 and 4 in UK-DALE, apparent power was recorded on aggregate level, whereas active power was considered on appliance level only. As our definition of NAR demands for the same AC power type on aggregate and submeter level, no such ratio could be computed in those cases. The same applies to all sites of the REDD (Kolter and Johnson 2011) dataset, according to the NILMTK dataset converter[1]. For this reason, REDD has not been considered in this study.

Evaluation setup

To gain a comprehensive understanding of the impact of noise on the disaggregation performance of algorithms, we selected three households with ascending levels of residual noise: household 2 of the ECO dataset (Beckel et al. 2014) with a NAR of 5.9%, household 5 of the UK-DALE dataset (Kelly and Knottenbelt 2015a) with a NAR 27.5%, and household 2 of the REFIT dataset (Murray et al. 2017) with a NAR of 65.1%. This way, we incorporate one instance each for disaggregation problems with low, moderate, and high noise levels. We selected five electrical appliances for every household considering a wide range of appliance types. We extracted 244 days for ECO, 82 days for UK-DALE and 275 days for REFIT while applying a sampling interval of 10 s. Table 2 provides further information on training and test sets. The amount of data used per household was governed by availability in the case of ECO and UK-DALE, as can be learned from Table 1. We split datasets into training set, validation set, and test set. This splitting was applied to all three households. We considered the smart meter signal as present in datasets and obtained

[1] https://github.com/nilmtk/nilmtk/tree/master/nilmtk/dataset_converters/redd/metadata

the denoised version of the aggregate by superposition of the individual appliance signals following:

$$y_t = \sum_{i=1}^{M} x_t^{(i)} \tag{3}$$

It should be noted that while deriving the denoised aggregate of a household, we considered all appliance signals available in the respective dataset. For instance, the denoised aggregate in the case of UK-DALE's house 5 is found by superposition of 24 appliance signals, as can be learned from Table 1.

For experimental evaluations, we utilize the latest version of NILMTK. The toolkit integrates several basic benchmark algorithms as well as load disaggregation algorithms based on Deep Neural Networks (DNN). In the course of experiments, we consider the traditional CO approach and two approaches based on DNNs:

- The *Combinatorial Optimization (CO)* algorithm, introduced in Hart (1992), has been used repeatedly in literature to serve as baseline (Batra et al. 2019). The CO algorithm estimates the power demand of appliances and their operational mode. Similar to the Knapsack problem (Rodriguez-Silva and Makonin 2019), estimation is performed by finding the combination of concurrently active appliances that minimizes the difference between aggregate signal and the sum of power demands.
- *Recurrent Neural Networks* are a subclass of neural networks that have been developed to process time series and related sequential data (Di Pietro and Hager 2019). First proposed for NILM in Kelly and Knottenbelt (2015b), we employ the implementation presented in Krystalakos et al. (2018), which incorporates Long Short-Term Memory (LSTM) cells. Provided a sequence of aggregate readings as input, the RNN estimates the power consumption of the electrical appliance it was trained to detect for each newly observed input sample.
- The *Sequence-to-point (S2P)* technique, relying on convolutional neural networks, follows a sliding window approach in which the network predicts the midpoint element of an output time window based on an input sequence consisting of aggregate power readings (Zhang et al. 2018). The basic idea behind this method is to implement a non-linear regression between input window and midpoint element, which has been applied successfully for speech and image processing (van den Oord et al. 2016). In a recent benchmarking study of NILM approaches, S2P was observed to be amongst the most advanced disaggregation techniques at that time (Reinhardt and Klemenjak 2020).

While the CO approach does not need to be parametrized, we set the number of training epochs to 25 during training of neural networks. Further, we employ an input sequence length of 49 for LSTM inspired by Krystalakos et al. (2018) and 99 for S2P as suggested in Batra et al. (2019).

In this study, we utilize two error metrics to assess the performance of load disaggregation algorithms. The first is a well-known, common metric used in signal processing, the Mean Absolute Error (MAE), defined as:

$$\mathrm{MAE}^{(i)} = \frac{1}{T} \cdot \sum_{t=1}^{T} |\hat{x}_t^{(i)} - x_t^{(i)}| \tag{4}$$

where x_t is the actual power consumption, $\hat{x}_t$ the estimated power consumption, and T represents the number of samples. The best possible value is zero and, as we estimate the power consumption of appliances, it is measured in Watts. As second metric, we incorporate a metric defined by NILM scholars in Kolter and Jaakkola (2012), the Normalized Disaggregation Error (NDE), defined as:

$$\mathrm{NDE}^{(i)} = \sqrt{\frac{\sum_{t=1}^{T}\left(\hat{x}_t^{(i)} - x_t^{(i)}\right)^2}{\sum_{t=1}^{T}\left(x_t^{(i)}\right)^2}} \tag{5}$$

In contrast to the MAE, the NDE represents a dimensionless metric and, more importantly, the NDE belongs to the class of normalized metrics. This allows for fair comparisons of disaggregation performance between appliance types (Klemenjak et al. 2020).

Results

We summarize the outcome of our investigations in Table 3 for the MAE and Table 4 with regard to the NDE. For several appliances per household, we compare the disaggregation performance of CO, LSTM, and S2P when applied to the real-world aggregate signal, denoted as *Real*, and the denoised aggregate signal *Den*, respectively.

In virtually all cases, we observe a strong tendency towards disaggregation algorithms providing better performance on denoised aggregate signals. In the context of error metrics such as MAE and NDE this means that the error observed on the real aggregate is larger than the error on the denoised aggregate. This holds true for almost all households and appliances considered, though some exceptions were identified: we spot a few cases in Table 3, namely the fridge and kettle in ECO as well as the dishwasher in UK-DALE showing the opposite trend for the CO algorithm. Same applies to all fridges with regard to the NDE metric, as Table 4 reports. It should be pointed out that in those cases, the performance of CO on the real-world and denoised aggregate signal shows a considerable gap when compared to LSTM and S2P. Therefore and because of CO being a trivial benchmarking algorithm, we claim that these cases can be neglected.

Table 3 Mean absolute error (MAE) in Watts for real and denoised testing

		CO		LSTM		S2P	
	Appliance	**Real**	**Den**	**Real**	**Den**	**Real**	**Den**
ECO (2)	audio system	37.7	32.3	6.4	5.6	6.8	5.9
NAR = 5.9%	dishwasher	43.1	40.2	7.5	3.9	5.8	3.6
	fridge	41.8	49.8	9.5	11.7	7.5	8.5
	kettle	17.3	42.7	4.2	2.5	3.2	1.3
	lamp	62.2	47.1	28.9	16.4	28.4	16.5
UK-DALE (5)	dishwasher	87.6	95.1	12.2	3.6	8.3	4.1
NAR = 27.5%	electric oven	50.7	39.2	20.7	9.0	17.4	9.1
	electric stove	131.3	40.1	7.3	6.1	6.4	4.4
	fridge	161.4	141.4	25.7	21.1	20.6	16.4
	washing machine	74.8	51.4	28.6	14.8	17.3	13.7
REFIT (2)	dishwasher	96.0	41.3	31.4	9.0	25.3	8.5
NAR = 65.1%	fridge	57.8	22.8	23.0	10.4	23.9	12.4
	kettle	79.1	9.2	9.8	3.3	9.7	3.2
	microwave	70.8	46.6	2.7	1.5	2.8	1.1
	washing machine	101.5	41.0	21.6	12.6	24.1	11.3

Table 4 Normalised disaggregation error (NDE) for real and denoised testing

		CO		LSTM		S2P	
	Appliance	**Real**	**Den**	**Real**	**Den**	**Real**	**Den**
ECO (2)	audio system	1.83	1.74	0.48	0.42	0.47	0.44
NAR = 5.9%	dishwasher	0.96	0.78	0.44	0.26	0.34	0.22
	fridge	1.8	2.07	0.45	0.5	0.38	0.36
	kettle	1.3	1.66	0.51	0.34	0.48	0.22
	lamp	1.27	1.18	0.74	0.5	0.74	0.52
UK-DALE (5)	dishwasher	1.44	1.55	0.76	0.39	0.61	0.34
NAR = 27.5%	electric oven	1.51	0.98	0.66	0.38	0.53	0.33
	electric stove	2.9	1.82	0.66	0.58	0.6	0.42
	fridge	3.16	3.24	0.64	0.6	0.55	0.46
	washing machine	1.47	1.16	0.63	0.35	0.42	0.32
REFIT (2)	dishwasher	1.06	0.74	0.56	0.19	0.48	0.18
NAR = 65.1%	fridge	1.4	1.81	0.7	0.48	0.68	0.48
	kettle	1.33	0.43	0.48	0.2	0.46	0.2
	microwave	4.54	3.84	0.85	0.45	0.84	0.36
	washing machine	2.04	1.42	0.82	0.5	0.75	0.45

As concerns LSTM and S2P, we identify a single contradictory observation, namely in the case of the fridge in ECO's household 2. In this particular case, we observed that testing on the real-world aggregate signal results in marginally better performance. One explanation for this could be the extremely low NAR in this scenario, 5.9%, and the fridge belonging to the category of appliances with a recurrent pattern (Reinhardt and Klemenjak 2020).

Having identified a clear tendency towards CO, LSTM, and S2P providing significantly better performance in the denoised signal case i.e. lower MAE and NDE, we draw our attention to the open question whether or not there exists a link between noise level and the magnitude of the performance gap between *Real* and *Den*. To investigate further in this, we define the performance gap to be the distance between the error on the real aggregate signal and the error observed signal when testing on the denoised aggregate signal:

$$\Delta\text{MAE} = \text{MAE}_{\text{real}} - \text{MAE}_{\text{denoised}} \tag{6}$$

$$\Delta\text{NDE} = \text{NDE}_{\text{real}} - \text{NDE}_{\text{denoised}} \tag{7}$$

We derive ΔMAE for the cases presented in Table 3 and illustrate an excerpt of found gaps in Fig. 3 for ECO, Fig. 4 for UK-DALE, and Fig. 5 for REFIT, where the focus of this discussion lies on the two approaches based on neural networks.

We observe clear gaps for both NILM approaches based on neural nets, LSTM and S2P. The illustrations show that neither approach seems to be resilient to noise. This is particularly interesting as approaches relying on LSTM cells as well as sequence-to-sequence learning have received increased interest lately (Reinhardt and Klemenjak 2020; Kaselimi et al. 2019; Kaselimi et al. 2020; Mauch and Yang 2015; Wang et al. 2019). Further, we identify higher performance gaps in test cases on REFIT's house 2 compared to house 5 of UK-DALE in this study. This is particularly apparent when comparing the performance gap for the dishwasher across households, where we measure a ΔMAE many times higher in case of REFIT. Also, we observe performance gaps twice as high for the fridge on REFIT

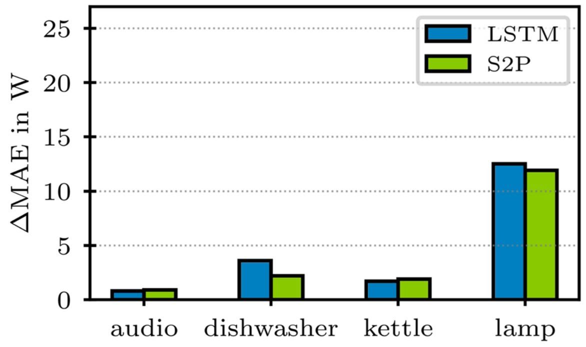

Fig. 3 Performance gap with regard to MAE for ECO house 2

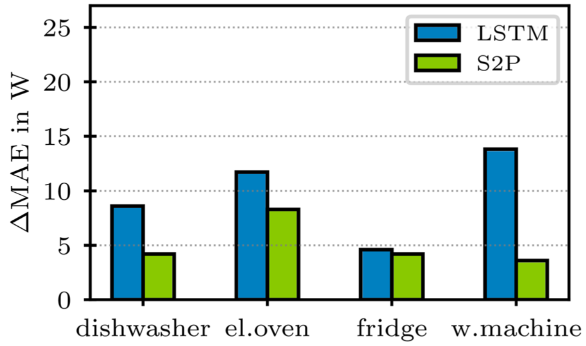

Fig. 4 Performance gap with regard to MAE for UK-DALE house 5

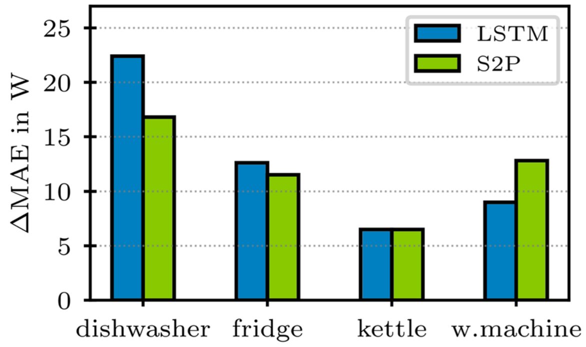

Fig. 5 Performance gap with regard to MAE for REFIT house 2

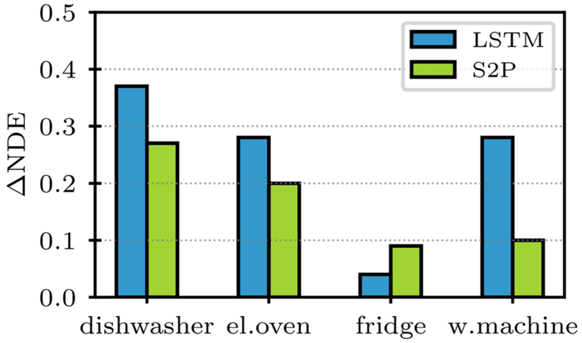

Fig. 6 Performance gap with regard to NDE for UK-DALE house 5

compared to UK-DALE. The only exception to this trend represents the case of LSTM for washing machines, where the performance gap of the LSTM network is smaller on REFIT than on UK-DALE.

Nevertheless, it should be stressed that comparisons based on not-normalized metrics can, but not have to be, misleading in some cases since two appliances of the same kind (i.e., two dishwashers) may differ significantly in terms of power consumption. Furthermore, metrics are designed to measure specific aspects of algorithms and hence, considering several metrics during performance evaluation results in a broader understanding of the capabilities of algorithms.

For these reasons, we also derived performance gaps with regard to NDE, ΔNDE, for the test cases presented in Table 4 and illustrate derived gaps in Fig. 6 for UK-DALE and Fig. 7 for REFIT.

In the case of fridges, we observe substantially lower performance gaps on UK-DALE for both networks. We suspect that is a result of the comparably high amount of noise in REFIT 2, disaggregating the real-world aggregate signal represents a bigger challenge than in the case of the denoised counterpart, especially when estimating the power consumption of low-power household appliances such as fridges.

Interestingly, not only we observe considerable performance gaps when estimating the power consumption of low-power appliances but also for appliances with moderate or high power consumption such as dishwashers and washing machines, as can be learned from Figs. 8 and 9. In both cases, UK-DALE and REFIT, we measure the highest ΔNDE in

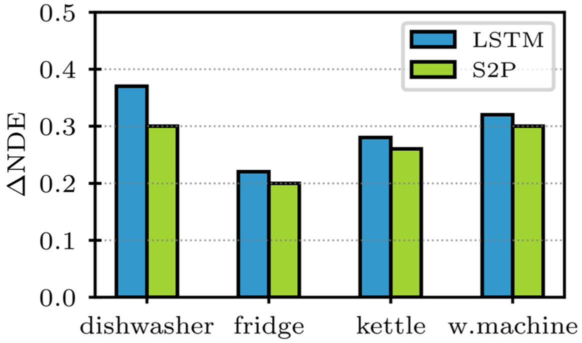

Fig. 7 Performance gap with regard to NDE for REFIT house 2

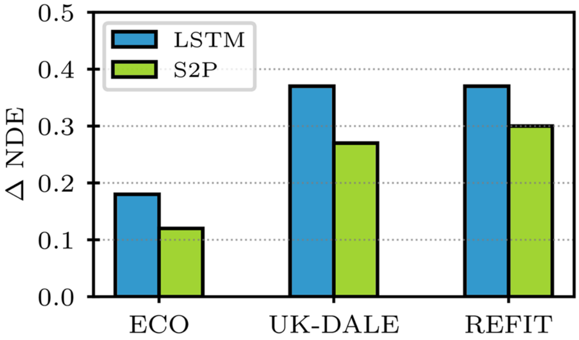

Fig. 8 Performance gap with regard to NDE for dishwashers

the case of the dishwasher. A comparison of performance gaps for dishwashers in Fig. 8 reveals that while we measure similar performance gaps in UK-DALE and REFIT, the performance gap in the case of ECO is significantly smaller. We hypothesize this is the result of the marginal noise level measured in house 2 of ECO. More importantly, we observe that also in cases of marginal noise levels, an apparent difference in terms of disaggregation error can be observed between real and denoised testing in this example.

A recent benchmarking study involving eight disaggregation algorithms found that S2P outperformed competing neural network architectures and concluded that S2P ranks amongst the most promising NILM approaches (Reinhardt and Klemenjak 2020). As concerns performance of NILM algorithms interpreted as disaggregation error between estimated power consumption and true power consumption of appliances, we find that S2P outperforms LSTM in 11 of 15 cases for the MAE metric and in 14 of 15 cases when the NDE metric is considered. Furthermore, in the vast majority of test runs, the S2P approach shows lower performance gaps than the network composed of LSTM cells in the sense of ΔMAE and ΔNDE.

Discussion

Insights obtained from testing on three households with considerably different NAR levels reveal that in the majority of test runs, testing on the denoised aggregate signal leads to substantially lower estimation errors and therefore, higher estimation accuracy. A few cases showing the contrary trend were observed but can be reasonably explained. As this

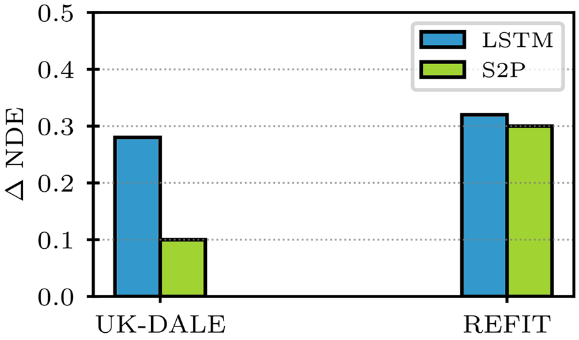

Fig. 9 Performance gap with regard to NDE for washing machines

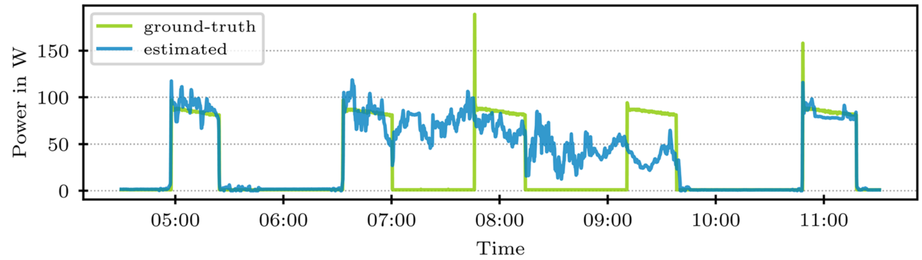

Fig. 10 An excerpt of estimates provided by S2P for the fridge in REFIT house 2 when applied to the real aggregate

apparent performance gap can be attributed to a variety of aspects, we suspect two of them having a decisive impact on this matter:

First, denoised aggregates are obtained by superposition of individual appliance signals. As such, they *contain fewer appliance activations and consumption patterns* than aggregates obtained from smart meters, respectively. Particularly when estimating the power consumption of low-power appliances, such activations have the potential to hinder load disaggregation algorithms from providing accurate power consumption estimates. Such cases were repeatedly observed during our studies on REFIT, where a NAR of 65.1% was measured. As depicted in Figs. 10 and 11, we detected several cases where concurrent operation of appliances with moderate or high power consumption (i.e. dishwasher, electric stove, or washing machine) resulted in significant deviations when estimating the power consumption of the fridge. Not only we observed such cases for the basic benchmarking algorithm CO but also for the advanced NILM approaches LSTM and S2P, which leads to the presumption that though having seen remarkable advances in the state of the art, at least a part of those algorithms may still be prone to noise levels in aggregate signals.

Second, we observe a substantially *higher number of false positive estimates* in predictions based on real-world aggregate signals than in estimates generated from denoised aggregate signals. False positives in this context mean that the NILM algorithms predicted the appliance to consume energy at times this was not the case. Such false positives impact the outcome of performance evaluations two-fold, as they increase the disaggregation error and decrease the estimation accuracy of NILM algorithms, respectively. We observed repeatedly that in the real-world case, the number of false-positive estimates

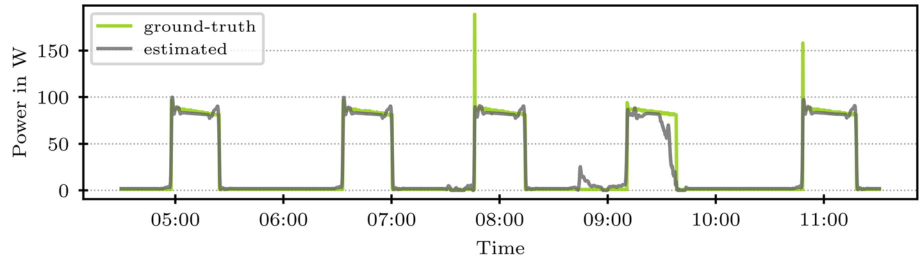

Fig. 11 An excerpt of estimates provided by S2P for the fridge in REFIT house 2 when applied to the denoised aggregate

is considerably higher than in the denoised case. We presume that those false positives are the result of algorithms confusing appliances with similar power consumption levels.

Based on the insights gained in this study, we can, however, not confirm a clear link between noise level, measured in NAR, and the magnitude of the performance gap between testing on real and denoised aggregates. We suspect this is due to the fact that every load disaggregation problem bears individual challenges to load disaggregation algorithms, making a comparison between moderate and high noise levels cumbersome. Though such a positive correlation between noise level and the magnitude of the performance gap could not be confirmed by our evaluation, we demonstrated that it has to be expected that testing on denoised aggregates results in lower disaggregation errors in the majority of test runs. Yet, we would like to stress the need for further investigation into the complexity of load disaggregation problems.

Conclusions

Motivated by the use of both, real and denoised aggregates in the evaluation of NILM algorithms in related work, we have investigated the performance gap observed between artificial sums of individual signals and signals obtained from real power meters. First, we utilized a noise measure, the noise-aggregate ratio NAR, to determine the noise level of real-world aggregate signals found in energy datasets. We find that noise levels vary substantially between households. We give insights on the experimental setup employed in our studies, comprising one basic and two more advanced NILM algorithms applied to data from three households with ascending noise levels. Our results show that a significant performance gap between the real and the denoised signal testing case can be identified in virtually all evaluation runs, provided a sufficiently high noise-aggregate ratio. Though some exceptions were observed, those cases can be well explained. Hence, we claim that testing on denoised aggregate signals can lead to a distorted image of the actual capabilities of load disaggregation algorithms in some cases, and ideally, its application should be well-considered when developing algorithms for real-world settings.

Abbreviations

AFAMAP: Additive factorial approximate maximum a posteriori; CO: Combinatorial optimization; DNN: Deep neural network; ECO: Electricity consumption and occupancy dataset; MAE: Mean absolute error; NAR: Noise aggregate ratio; NDE: Normalized disaggregation error; NILM: Non-intrusive load monitoring; NILMTK: Non-intrusive load monitoring toolkit; LSTM: Long short-term memory; REDD: Residential energy disaggregation dataset; S2P:Sequence-to-point learning

Acknowledgements

The authors acknowledge the financial support by the University of Klagenfurt.

Authors' contributions

CK designed and conducted all studies presented in this paper and wrote main parts of this manuscript. SM supervised the topic during CK's visit at Simon Fraser University. WE contributed to discussions and text, and supervised the topic at University of Klagenfurt. All authors read, edited, and approved the manuscript.

Funding

Not applicable.

Availability of data and materials

All datasets used in the course of our experiments are publicly available and can be obtained from the respective project homepage. NILMTK, the toolkit employed to conduct studies, can be downloaded from Github: https://github.com/nilmtk/nilmtk

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Institute of Networked and Embedded Systems, University of Klagenfurt, Klagenfurt, Austria. [2] School of Engineering Science, Simon Fraser University, BC V5A 1S6, Canada.

Received: 3 October 2020 Accepted: 11 January 2021
Published online: 04 March 2021

References

Anderson K, Ocneanu A, Benitez D, Carlson D, Rowe A, Berges M (2012) BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. In: Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD). ACM, Beijing. pp 1–5

Batra N, Gulati M, Singh A, Srivastava MB (2013) It's different: Insights into home energy consumption in India. In: Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings. pp 1–8

Batra N, Kelly J, Parson O, Dutta H, Knottenbelt W, Rogers A, Singh A, Srivastava M (2014b) NILMTK: an open source toolkit for non-intrusive load monitoring. In: Proceedings of the 5th International Conference on Future Energy Systems. pp 265–276

Batra N, Kukunuri R, Pandey A, Malakar R, Kumar R, Krystalakos O, Zhong M, Meira P, Parson O (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. pp 193–202

Batra N, Parson O, Berges M, Singh A, Rogers A (2014a) A comparison of non-intrusive load monitoring methods for commercial and residential buildings. arXiv preprint arXiv:1408.6595:1–11

Beckel C, Kleiminger W, Cicchetti R, Staake T, Santini S (2014) The ECO data set and the performance of non-intrusive load monitoring algorithms. In: Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings. pp 80–89

Bonfigli R, Felicetti A, Principi E, Fagiani M, Squartini S, Piazza F (2018) Denoising autoencoders for non-intrusive load monitoring: improvements and comparative evaluation. Energy and Build 158:1461–1474

Bonfigli R, Principi E, Fagiani M, Severini M, Squartini S, Piazza F (2017) Non-intrusive load monitoring by using active and reactive power in additive Factorial Hidden Markov Models. Appl Energy 208:1590–1607

Di Pietro R, Hager G (2019) Handbook of medical image computing and computer assisted intervention. Chapter 21:503–519

Gopinath R, Kumar M, Joshua CPC, Srinivas K (2020) Energy management using non-intrusive load monitoring techniques-State-of-the-art and future research directions. Sust Cities Soc 62:102411

Hart GW (1992) Nonintrusive appliance load monitoring. Proc IEEE 80(12):1870-91

Kaselimi M, Doulamis N, Doulamis A, Voulodimos A, Protopapadakis E (2019) Bayesian-optimized bidirectional LSTM regression model for non-intrusive load monitoring. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Brighton. pp 2747–2751

Kaselimi M, Doulamis N, Voulodimos A, Protopapadakis E, Doulamis A (2020) Context aware energy disaggregation using adaptive bidirectional LSTM models. IEEE Trans Smart Grid 11(4):3054–67

Kelly J, Knottenbelt W (2015a) The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Sci Data 2(1):1–14

Kelly, J, Knottenbelt W (2015b) 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 (BuildSys). pp 55–64

Klemenjak C, Makonin S, Elmenreich W (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). pp 1–5

Kolter JZ, Jaakkola T (2012) Approximate inference in additive factorial hmms with application to energy disaggregation. In: Artificial Intelligence and Statistics. pp 1472–1482

Kolter JZ, Johnson MJ (2011) Redd: A public data set for energy disaggregation research. In: Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA Vol. 25. pp 59–62

Krystalakos O, Nalmpantis C, Vrakas D (2018) Sliding window approach for online energy disaggregation using artificial neural networks. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence (SETN). pp 1–6

Makonin S, Ellert B, Bajic IV, Popowich F (2016) Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. Sci Data 3(160037):1–12

Makonin S, Popowich F (2015) Nonintrusive load monitoring (NILM) performance evaluation. Energy Efficiency 8(4):809–814

Makonin S, Popowich F, Bajić IV, Gill B, Bartram L (2015) Exploiting HMM sparsity to perform online real-time nonintrusive load monitoring. IEEE Trans Smart Grid 7(6):2575–2585

Mauch L, Yang B (2015) A new approach for supervised power disaggregation by using a deep recurrent LSTM network. In: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, Orlando. pp 63–67

Monacchi A, Egarter D, Elmenreich W, D'Alessandro S, Tonello AM (2014) GREEND: An energy consumption dataset of households in Italy and Austria. In: 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm). pp 511–516

Murray D, Stankovic L, Stankovic V (2017) An electrical load measurements dataset of united kingdom households from a two-year longitudinal study. Sci Data 4(1):1–12

Pereira L, Nunes N (2018) Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools–A review. Wiley Interdiscip Rev Data Min Knowl Disc 8(6):1265

Reinhardt A, Baumann P, Burgstahler D, Hollick M, Chonov H, Werner M, Steinmetz R (2012) On the accuracy of appliance identification based on distributed load metering data. In: 2012 Sustainable Internet and ICT for Sustainability (SustainIT). IEEE, Pisa. pp 1–9

Reinhardt A, Klemenjak C (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. Association for Computing Machinery, New York, NY, USA. pp 167–177

Rodriguez-Silva A, Makonin S (2019) Universal Non-Intrusive Load Monitoring (UNILM) Using Filter Pipelines, Probabilistic Knapsack, and Labelled Partition Maps. In: 2019 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC). pp 1–6

Salem H, Sayed-Mouchaweh M, Tagina M (2020) A Review on Non-intrusive Load Monitoring Approaches Based on Machine Learning. Springer, Cham

van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: A generative model for raw audio. In: 9th ISCA Speech Synthesis Workshop. pp 125–125

Wang K, Zhong H, Yu N, Xia Q (2019) Nonintrusive load monitoring based on sequence-to-sequence model with attention mechanism. In: Zhongguo Dianji Gongcheng Xuebao/Proceedings of the Chinese Society of Electrical Engineering Vol. 39. pp 75–83

Wittmann FM, López JC, Rider MJ (2018) Nonintrusive load monitoring algorithm using mixed-integer linear programming. IEEE Trans Consum Electron 64(2):180–187

Zhang C, Zhong M, Wang Z, Goddard N, Sutton C (2018) Sequence-to-point learning with neural networks for non-intrusive load monitoring. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI). pp 2604–2611

Zhao B, He K, Stankovic L, Stankovic V (2018) Improving event-based non-intrusive load monitoring using graph signal processing. IEEE Access 6:53944–53959

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com