Outlier Detection in Energy Datasets

An Honors Paper for the Department of Computer Science
Stephen Crawford

Bowdoin College, 2022

Abstract

In the past decade, numerous datasets have been released with the explicit goal of furthering non-intrusive load monitoring research (NILM). NILM is an energy measurement strategy that seeks to disaggregate building-scale loads. Disaggregation attempts to turn the energy consumption of a building into its constituent appliances. NILM algorithms require representative real-world measurements which has led institutions to publish and share their own datasets. NILM algorithms are designed, trained, and tested using the data presented in a small number of these NILM datasets. Many of the datasets contain arbitrarily selected devices. Likewise, the datasets themselves report aggregate load information from building(s) which are similarly selected arbitrarily. This raises the question of the representativeness of the datasets themselves as well as the algorithms based on their reports. One way to judge the representativeness of NILM datasets is to look for the presence of outliers in these datasets. This paper presents a novel method of identifying outlier devices from NILM datasets. With this identification process, it becomes possible to mitigate and measure the impact of outliers. This represents an important consideration to the long-term deployment of NILM algorithms.

Acknowledgements

This work would not be possible without the steady guidance of Professor Sean Barker. His wisdom and expertise were fundamental to the success of this research and he gave countless hours to the improvement of this manuscript. Further, this work would not have been possible without the years of excellent education from the Bowdoin College Department of Computer Science.

Contents

1 Introduction **1**
1.1 NILM Background 1
1.2 Small Sample Size 2
1.3 Selection Process 3
1.4 Appliance Description 4

2 Related Works **6**
2.1 NILM Datasets 6
2.2 Clustering Research 7
2.2.1 Short-Term Load Forecasting 7
2.2.2 Energy Disaggreation 8
2.2.3 Hybrid Disaggregation 9

3 Methods **10**
3.1 Algorithm Overview 10
3.2 soft-Dynamic Time Warping 12
3.3 Timeseries K-Means 12

4 Discussion **14**
4.1 Dataset Overview and Experimental Parameters 14
4.2 Outlier Detection Results 15
4.3 Sample Comparison 17

5 Conclusion **21**
5.1 Contributions 21
5.2 Future Work 21

List of Figures

1 The methodology of the clustering algorithm. 11
2 The results of clustering the fridge for 12 hours. The red line represents the Barycenter Average of the cluster. The blue lines represent cluster constituents. 14
3 A comparison of the identified fridge outlier (black) versus a random subset of other fridges. 18
4 A comparison of the identified furnace outlier (black) versus a random subset of other furnaces. 19
5 A comparison of another outlier fridge. Outlier (black) vs. random subset of other fridges. 20

1 Introduction

1.1 NILM Background

Non-Intrusive Load Monitoring (NILM) is the process of disaggregating building-scale energy consumption readings in order to identify the energy use of specific appliances. The primary motivator behind NILM research is evidence that suggests that with more accurate forecasting and analysis of smart meter data, companies can provide actionable feedback to consumers who can, in turn, improve energy efficiency by up to 15% [1]. It has also been suggested that it may be possible for energy providers to tailor a customer's service plan to best support their own needs while reducing overall grid load and better meeting environmental concerns [2], [3]. Originally proposed by George Hart in 1985, NILM has seen a resurgence in popularity coinciding with the increasing availability and use of smart meter technology. Smart meters have seen deployment around the world in both residential and commercial buildings [4].

With the resurgence of NILM research, numerous dedicated datasets have been published, Table 1 details the basic information for the most popular of these datasets. Alongside their stated purpose in NILM research, these datasets have been adopted for numerous other applications. Some of the most notable lines of inquiry, include consumption forecasting [5], [6], demand-side management [7]–[9], consumer behavior analysis [10]–[12], and appliance anomaly detection [13].

Across the many areas where NILM datasets are used, there lies the common issue of representativeness. Broadly, research is only valuable to the extent that it can be generalized. This reality is even more important when the research concerns NILM or other energy or behavior-related disciplines. For applications involving NILM, behavior prediction, or energy forecasting to be effective, the algorithms being used need to be deployable on a broad scale. However, several issues persist within many of the most commonly cited datasets. These issues bring the representativeness of the datasets into question. Specifically, most major datasets suffer from some combination of issues within three categories: 1) a limited number of samples; 2) a seemingly arbitrary selection process; 3) a lack of description as to the specifics of appliances measured.

Dataset	Houses	Type
REDD (2011)	6	Non-event based
BLUED (2012)	1	Event-based
SMART* (2012)	3	Both
HES (2012)	251	Non-event based
AMPds (2013)	1	Non-event based
iAWE (2013)	1	Non-event based
UK-DALE (2014)	5	Non-event based
ECO (2014)	6	Non-event based
GREEND (2014)	9	Non-event based
SustData (2014)	50	Non-event based[1]
Dataport (2014)	722	Non-event based
DRED (2015)	1	Non-event based
PLAID (2017)	64	Both
ENERTALK (2019)	22	Non-event based
MORED (2020)	13	Both

Table 1: The most popular NILM datasets with the number of homes they contain and the measurement type. [1] An extension of the dataset includes event-based readings.

1.2 Small Sample Size

While there is no inherent issue in the release of small-scale datasets, scientific conclusions based on small datasets are more likely to suffer inaccuracies. NILM datasets often contain measurements for only a small number of houses as a result of practical and financial concerns. While individual meters are not prohibitively expensive, the costs associated with the collection of large-scale datasets far exceed the typical research budget [14]. For example, the budget for SMART*, of around 3000 dollars, was only enough to cover the metering of three homes. This reality provides insight into why many relevant datasets contain fewer than 10 monitored homes. Fortunately, of the three potential hurdles for NILM datasets, sample count is the one area that has been most readily addressed. While older datasets with few monitored houses remain in use, recently released datasets such as SustData and Dataport are magnitudes larger than their predecessors (Table 1). Unfortunately, the other two concerns—of sample selection process and data specificity—remain largely unaddressed.

1.3 Selection Process

The selection criteria for buildings monitored for the creation of NILM datasets is important for numerous reasons. First, with little to no details being offered about the selection process, studies prevent any form of true comparison. When datasets are released without specifying the data sources, there is no method for researchers outside of the original publishers to corroborate findings. While the studies based on the released datasets may be reproducible, the datasets themselves must be taken at face value. Some datasets provide a detailed description of monitoring equipment and the metering setup itself, but there is no guarantee that the use of the same protocols will yield comparable data when deployed in an entirely different home. It is not uncommon for datasets to contain some buildings where every appliance is sub-metered and others where only some—or maybe no—appliances are sub-metered. This can cause issues for future works relying on the data. In the best-case scenario, future works must make some chain of assumptions to make use of the buildings with less data; in the worst cases, researchers are forced to exclude certain buildings and further decrease the size of already small datasets. Though there are valid privacy concerns which preclude the release of identifying information, broad information such as economic standing, home size, and construction/renovation year at least provide a rough picture of the participants. It is unlikely that older appliances–perhaps more common in older homes–operate in the same manner as newer appliances. Likewise, it cannot be assumed that participants who fall into vastly different socioeconomic categories own appliances of similar make or energy efficiency. Candidly, it is not obvious that those *within* the same socioeconomic category can be expected to own comparable devices.

To this point, a second issue arising from the lack of a specified selection process is behavioral profiles associated with different populations. When there is an arbitrary selection process, generalizing to a broader population becomes difficult. For example, the SMART* dataset notes that the monitored homes were those of graduate students participating in research with the investigators [14]. Graduate students are unlikely to have similar usage patterns to a family where adults work from home or that of retirees. While the duration of usage is often presumed insignificant to the representativeness of sub-metered loads, there is no guarantee that certain appliances do not perform differently under more or less frequent operation. It is also worth recalling that NILM datasets are also used for behavioral monitoring where differences in occupancy are likely to have an even greater impact.

Finally, geographic differences are likely to play a significant factor in appliance load and use. Consider that countries have separate standardized voltages (United Kingdom (230 V) and the United States (120 V)). There may be an underlying difference between appliances designed for different locations. Even datasets from within the same country or region are likely to show a great deal of variance with regard to their operation. Datasets collected in the United States for instance could contain data collected from diverse climates such as the cold and dry New England region, the humid and hot Gulf Coast, the arid Southern Midwest, or the perpetually rainy Northwest. In each of these environments, even participants who may otherwise have similar lifestyles and occupancy patterns are likely to have vastly different appliance usage. This is particularly true for HVAC units.

1.4 Appliance Description

The final concern potentially hampering the utility of algorithms based on NILM datasets is the general lack of a description for the monitored appliances. The problems associated with not having a specific description of the appliances monitored in each building are significant. Generally, the issues involve many of the same assumptions discussed with regard to the selection process. More specifically, without detailing the make, model, and condition of monitored appliances, it is unlikely that research based on NILM datasets can mitigate their own assumptions. The fewer details provided about the monitored appliances, the broader the hypothesis space.

As was seen with the absence of a selection criteria, failure to specify appliance details makes any attempts at reproduction difficult. As was mentioned, a participant living in an older home is unlikely to have the same type of refrigerator as someone living in a recently built apartment building. As appliance technology has evolved, numerous factors have contributed to an expected change in their load consumption. Generally, it would be expected that energy efficiency will have increased alongside broad efforts to conserve electricity. Energy efficiency ratings such as the United States Government-run Energy Star program have incentivized the creation, promotion, and use of appliances that meet certain levels of energy efficiency [15]. While the Energy Star programs' specifications are not particularly rigorous, there is likely a great deal of variance in the load patterns of appliances that fail to meet Energy Star standards. An appliance may be just below the requirement threshold, or it could be a massive power sink. Researchers behind the first public NILM dataset note that: "generalization across homes and device categories make disaggregation

much more complicated" [16].

Furthermore, even if a selection process was meticulously detailed with a dataset's release—including one that went so far as to denote the exact home—without specific appliance information, it would be difficult for researchers to verify findings or collect additional data. Without device information, the researchers could not be certain that a given appliance was the same one initially monitored. Most likely, they would have to ask the occupant whether it was the same appliance which is an unscientific means of verification at best. Such an approach would be vulnerable to human error and prove impossible should the original occupants have moved.

Concerns about the specific appliances measured are not only limited to cases where datasets may be using inefficient or outdated appliances. Even if all other factors are similar, there is no guarantee that two appliances of the same type perform the same way. It is reasonable to imagine a scenario where one washing machine performs significantly differently than another even when accounting for major factors such as which year they were made and where they are used. The obvious solution to this issue would be to set a specific make and model for the appliances used and only collect data from these devices. The issue with this solution is that it effectively counteracts the progress made in expanding the number of samples. If the goal is ultimately for datasets to be more representative, then restricting the domain space to a small variety of specific appliances is counterproductive. Instead of making findings more applicable to unseen cases, this would likely cause severe overfitting [17].

Finally, there is a level of expected variance amongst samples even with all other factors accounted for. Even if a dataset were to choose to focus on only a single make and model of appliance (and also accounted for the numerous other considerations listed thus far), there is still expected variation between one appliance and another. Because there is no practical solution to the issue of selection criteria or appliance description, it is important to be able to determine whether a dataset is representative. One method of doing so which not require the full-scale deployment of an algorithm is analysis of a dataset's outliers. For this purpose, this paper presents a novel method of detecting outliers within NILM datasets. Using this strategy, it is feasible to quantitatively analyze the impact of outliers on algorithms based on the datasets.

The remainder of this paper is structured as follows: Section 2.1 describes a selection of the most popular NILM datasets. Section 2.2 discusses related clustering

work focusing on NILM datasets. Section 3 details the methods of analysis used. Section 4 discusses the results of the empirical evaluation of the algorithm. Finally, Section 5 offers insight into future applications of the proposed outlier detection algorithm.

2 Related Works

2.1 NILM Datasets

The purpose of this section is to provide a cursory glance at a few of the most popular NILM datasets. For each dataset, the location, notable features, and time span are noted. Explicit values are provided for the number of citing works as well as the number of times notable citing works (excluding other NILM datasets) have themselves been cited. The purpose of this overview is to demonstrate the prominence of NILM datasets within two degrees and the associated possibility for unrepresentative findings.

1. **REDD**
   The Reference Energy Disaggregation Data Set (REDD) was the first published NILM dataset. Before its release, NILM research was conducted using proprietary datasets which all but assured that findings were irreproducible. The researchers behind REDD sought to rectify this and established REDD as an open-source NILM dataset accessible to anyone looking to analyze energy consumption. The REDD dataset contains data for 6 homes. For each home, the researchers measured: "the whole home electricity signal recorded at a high frequency (15kHz); up to 24 individual circuits in the home, each labeled with its category of appliance or appliances, recorded at 0.5 Hz; [and] up to 20 plug-level monitors in the home, recorded at 1 Hz..." [16] The dataset spans a period of several months and the homes are all located in the Greater Boston area of Massachusetts. The REDD dataset has been cited 1319 times. Notable publications based on the REDD dataset include: [18] with 666 citations; [19] with 526 citations; [20] with 370 citations.

2. **SMART***
   The SMART* dataset publication goes into an extensive description of the three homes its researchers measured, however, does not list the explicit appliances used. SMART* offers event tracking for numerous appliances and

devices including wall switches and thermostats. The dataset contains both real and reactive power consumption measurements. The publication makes explicit mention of the size of each home as well as describes the relocation of one of the home's wind turbine. The homes are stated to be in Western Massachusetts. The publication associated with SMART* has been cited 450 times[14]. Some notable citing works are: [21] with 241 citations; [22] 184 citations; and [23] with 164 citations.

3. **UK-DALE**
   The UK Domestic Appliance-Level Electricity (UK-DALE) dataset contains data from 5 homes collected over a period of up to 655 days. The researchers state that "the subjects were either MSc students or Ph.D. students at Imperial College. The subjects chose to do a research project with the authors... The upper bound on the number of houses we could record from was set by a combination of a limited financial budget, limited time to assemble the metering hardware, and a limited number of [volunteer students]"[24]. Numerous works such as [25] with 228 citations, [26] with 210 citations, and [27] with 161 citations make use of the dataset in their research. UK-DALE's associated publication has been cited 565 times.

2.2 Clustering Research

In order to determine the novelty of this work, comparison can also be made between this research and other clustering work which focuses on NILM datasets. In considering related works, there are three main categories that share similar methodology to this study.

2.2.1 Short-Term Load Forecasting

The first category of research using clustering on NILM datasets is known as Short-Term Load Forecasting (STLF). While STLF itself has been heavily researched, the use of NILM data with STLF is relatively new. STLF research is concerned with the supply-demand balance of consumers and demand-side managers. Broadly, the goal of the research is to develop algorithms that can effectively predict how much energy a given residence is likely to require

at any given point in time. By accurately and efficiently forecasting energy needs, energy providers can optimize their generation and delivery strategies. As NILM datasets have become more common, STLF researchers have become interested in using them to assist their research. Examples of STLF work which incorporate clustering data from NILM datasets include: [28], [29], [30], and [31].

Of these works, a typical example of NILM-based STLF is the work completed by Dinesh et al. [30]. In this work, the researchers used aggregate energy readings from NILM datasets such as REDD and AMPds2 (the revised version of AMPds). The researchers then decomposed the aggregate data into individual appliances before forecasting the individual future loads. They finally recombined the future loads to calculate a building-scale prediction. In the case of Dinesh et al., spectral clustering was used in order to predict which appliances would be on at any given time. Their spectral clustering approach includes several steps which are beyond the scope of this paper but the internal process involved is an application of K-means clustering to the spectral representations of an appliance's correlation of operation with the other monitored appliances [30]. As should be obvious, while STLF based on clustering NILM datasets incorporates many similar techniques to this paper, the goals of the research are quite different. Where STLF work seeks to predict the amount of energy a given appliance and home may require, this work seeks to identify outliers that may prevent algorithms based on NILM datasets from generalizing.

2.2.2 Energy Disaggreation

A second area of research involving clustering NILM datasets is the direct application of clustering methods to disaggregate the aggregate energy readings. This is likely the most obvious application of clustering of NILM datasets. In this line of inquiry, clustering is used to identify the states of different appliances. This information is then fed into a classification and load estimation system. Examples of this particular application of clustering to NILM data include: [32]; [33]; [34]; [35]; [36]; and [37].

A typical example of research focused on directly disaggregating NILM data via clustering is represented by Barsim et al. [33]. In their influential research, the authors use two clustering algorithms in order to disaggregate the

BLUED dataset and achieve a 92% disaggregation accuracy and a 98% clustering accuracy. As opposed to a traditional change-point detection method of determining transient appliance states, the authors utilize grid-based clustering for this purpose. The researchers note that this allows them to determine the exact time window of the transient states which in turn allows for the identification of appliance signatures from their transient behavior. Furthermore, the use of grid-based clustering provides a computationally efficient method of running the clustering-based event detection process. After identifying the transient states, features are extracted from each identified transient state.

In order to cluster events, the researchers cluster based on the features extracted from the transient states. This is done with mean-shift clustering which is a non-parametric clustering algorithm that has recently entered the NILM spotlight. By using mean-shift clustering, the number of appliances does not need to be known a priori. Furthermore, mean-shift clustering is independent of any distribution of appliances and has an implicit mode-seeking function. Finally, the researchers conclude their disaggregation by using a ground-state (the lowest steady-state) pairing process.

As with the case of NILM-based STLF, it should be obvious that while this work uses clustering on NILM datasets, the application of clustering is extremely different from this study. In fact, the work of Barsim et al. would likely benefit from knowing the extent to which BLUED and other NILM datasets are believed to generalize.

2.2.3 Hybrid Disaggregation

A third and final type of research that involves clustering NILM data is hybrid energy disaggregation approaches. This third category of NILM-focused clustering research has by far the most variety within it. Each publication generally takes quite different approaches to the use of clustering in the overall task of disaggregation. These techniques use a clustering algorithm to analyze or reason with features extracted from data by a different type of algorithm. For example, [38] makes use of clustering to identify steady-states while using a more traditional approach to identify transient-states. [39] however, clusters on features which have already been processed by regression trees. Examples of hybrid disaggregation approaches include [38], [39], and [40].

Due to the variety of research within this category, it is difficult to point to a single work as an exemplary piece. That being said, it should once again be obvious that these works are not particularly similar to the research conducted in this study. As was seen with the first two categories of NILM-focused clustering research, the works associated with hybrid energy disaggregation stand to benefit from the findings of this paper given the ultimate deployment of NILM algorithms being the works' implicit goal.

Ultimately, in all three cases of NILM-focused clustering research, the research is not overly similar to this paper's own work. Instead, the works are likely to benefit from the outlier detection strategy proposed by this work. Using the proposed algorithm, the researchers will be able to better assess the extent their algorithms can be deployed.

3 Methods

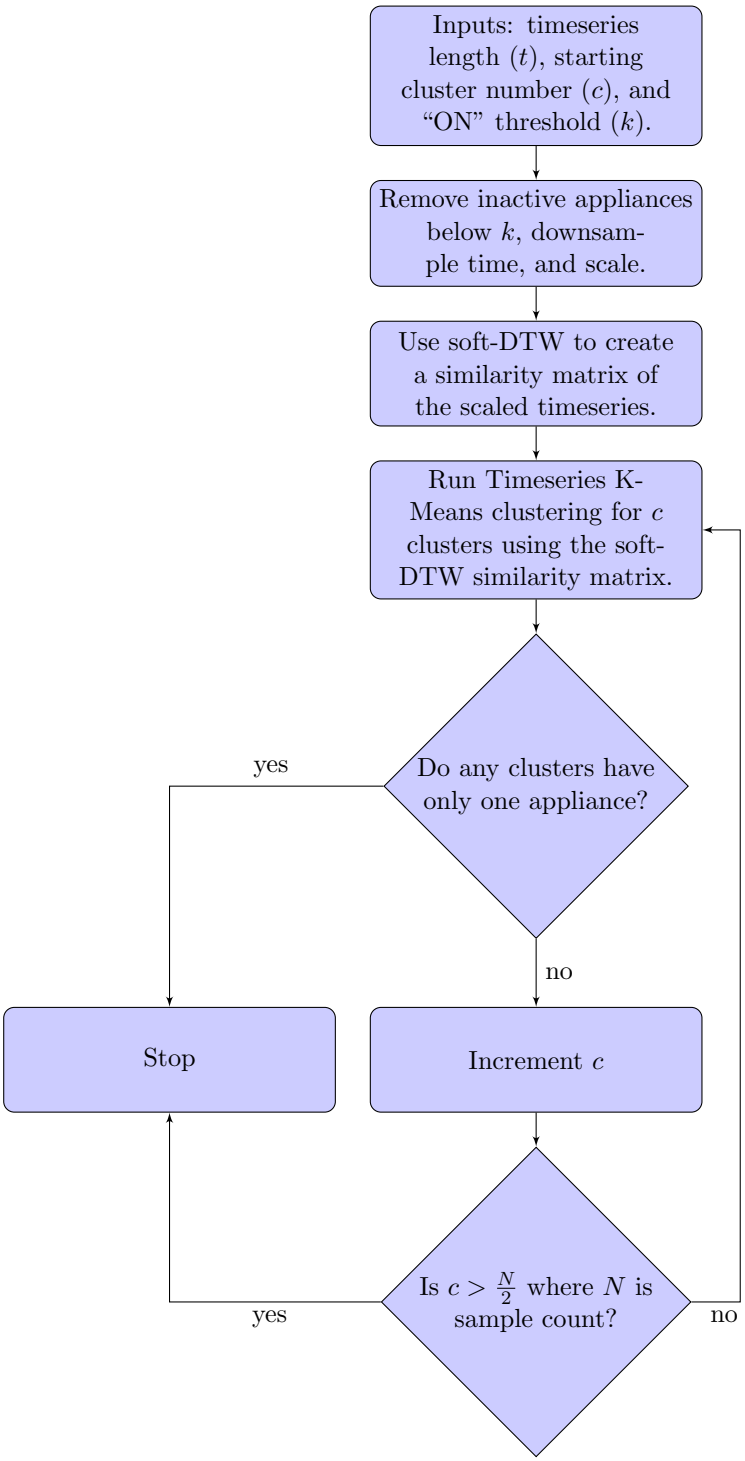
This section details the development of the outlier detection algorithm as well as the experimental framework. To reiterate the purpose of the algorithm: the algorithm needs to be able to analyze a NILM dataset and identify anomalous appliances within the different appliance categories. The algorithm is intended to be used by researchers working with NILM datasets so that they can measure the impact of outliers in the datasets on their final product.

3.1 Algorithm Overview

In order to provide a clear framework for the interweaving of the algorithmic foundations of the outlier detection process, please refer to Figure 1.

Having provided a prose description of the outlier identification process, it is now possible to further examine the specific components in detail. Specifically, the following subsections discuss the use of soft-Dynamic Time Warping for the calculation of a similarity matrix, Timeseries K-Means Clustering for the execution of the actual clustering sub-process, and the use of Barycenter Averaging for a calculation of the typical load-signature within each cluster.

Figure 1: The methodology of the clustering algorithm.



3.2 soft-Dynamic Time Warping

The first foundation of the outlier detection algorithm is the use of soft-Dynamic Time Warping (sDTW). sDTW is used to calculate the similarity matrix which, alongside Timeseries K-Means Clustering, is used by the algorithm to identify anomalous appliances. Dynamic Time Warping (DTW) is a method of finding the ideal alignment between two time-dependent series of values [41]. DTW uses dynamic programming to identify the minimized distances between the values in timeseries $S = s_1, s_2, \dots s_n$ and another timeseries $T = t_1, t_2, \dots t_n$ [42]. As illustrated by Berndt and Clifford, DTW identifies a "Warping Path" $W = w_1, w_2, ..., w_n$ such that when the weights of W are applied to the values of S and T, the distance between S and T is minimized. The distance to be minimized, δ, represents a hyperparameter. In this algorithm, the distance is simply the euclidean distance between the two timeseries after having been scaled. Using DTW, it is possible to create an $N \times N$ matrix where N is the number of instances of appliances. For example, if the dataset has 14 homes with *microwaves* then $N_{microwave} = 14$. For each entry in the $N \times N$ matrix, an optimized solution to the DTW problem is stored. The optimized solution stored in each matrix entry (i, j) is the $W_{(i,j)}$ weights which minimizes the sum of distances between points in the i^{th} and j^{th} timeseries. sDTW improves on the computational cost of true DTW by computing the soft-minimum of the optimization problem. As a result, the loss function is differentiable and can be computed in quadratic space and time complexity [43]. sDTW replaces the original minimum calculation within DTW with a soft-minimum calculation which is differentiable with the chain-rule, and which results in less noisy Barycenter Average calculations.

3.3 Timeseries K-Means

Using sDTW as the distance metric, it is possible to perform Timeseries K-Means clustering. As shown in Figure 1, the clustering process is used iteratively in order to detect anomalous appliances. For each iteration, the data is clustered and then the clusters themselves are analyzed according to the three end conditions listed below. Clustering was used as the method of outlier detection because of the variability of different appliance types. While it would be possible to define a normalized definition of an outlier, it is suspected that

different appliance types have intrinsically different thresholds for atypical behavior. That is, while an outlier defined for fridges may be a fridge that is less than 2% likely to be produced via kernel density estimation, a different appliance type may inherently be more or less susceptible to anomalous behavior. As such, the threshold for the likelihood of generating a sample that classifies it as an outlier may not be the same across categories. Clustering approaches for outlier identification face no such limitation and can thus be executed with the exact same hyperparameters for all appliance types. This is imperative when considering the algorithm is intended to be easy to deploy with little to no maintenance.

There are three end cases for the clustering algorithm as it attempts to identify an outlier appliance: 1) No outliers could be found for numerous iterations of the clustering process. It was considered reasonable to state that this method of outlier identification could not detect any outliers if the number of clusters was increased to more than $\frac{N}{2}$ where N is the number of active appliance samples in the appliance DataFrame. When the number of clusters grew over $\frac{N}{2}$, there was necessarily going to be at least two appliances which were in their own clusters so this result would not be indicative of anything meaningful. 2) A cluster was formed in under $\frac{N}{2}$ iterations which had only a single appliance would indicate that the most unusual or anomalous appliance had been found. 3) Multiple clusters were formed in under $\frac{N}{2}$ which had only a single appliance in them. This situation could arise in the case where a cluster of two appliances is split into two different clusters when the cluster count is incremented.

The implementation of the algorithm used in this research also involved visualization of the Barycenter Averaging of the clusters. This allowed for convenient visualization of the cluster constituents as well as the 'typical' scaled appliance measurement within that cluster. An example of this technique is shown in Figure 1. Barycenter Averaging calculates the center timeseries for each cluster output by the Timeseries K-Means Clustering and is commonly used in applications of DTW. Further, Barycenter Averaging is particularly valuable when analyzing NILM datasets as it provides a stand-in for the average appliance load pattern of each cluster.

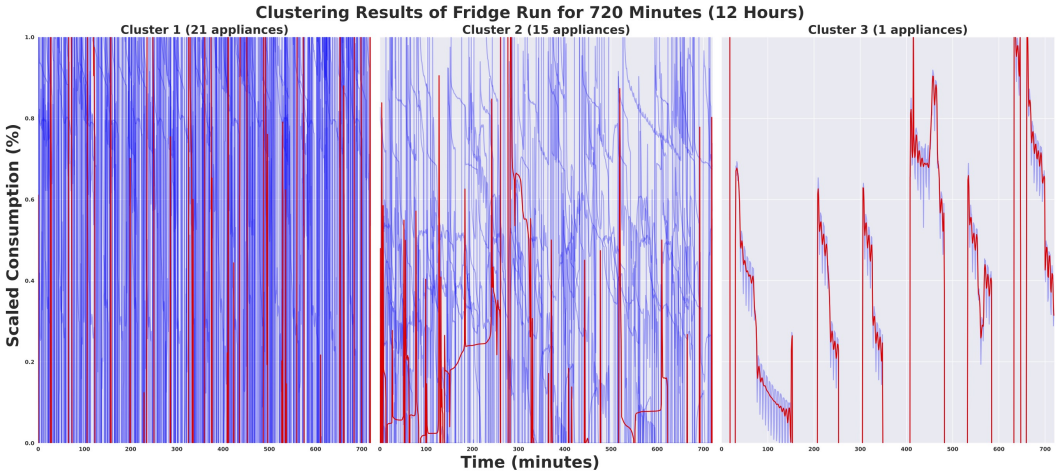


Figure 2: The results of clustering the fridge for 12 hours. The red line represents the Barycenter Average of the cluster. The blue lines represent cluster constituents.

4 Discussion

This section provides an overview of the process of empirical evaluation of the outlier detection algorithm. The section is broken into two subsections. Section 4.1 provides an overview of the dataset which was used during experimentation. It also offers a specific listing of the hyperparameter settings used for experimentation. Section 4.2 discusses the results of the preliminary clustering experimentation and outlier detection process. Finally. Section 4.3 details the results of experimental evaluation by analysis of the detected outliers through comparison to random samples in the dataset.

4.1 Dataset Overview and Experimental Parameters

The process of empirical evaluation of the outlier detection algorithm made use of a 50 home subset of the Pecan Street Dataport dataset. Using this subset, the outlier detection algorithm was used to cluster 12 different submetered devices and the house-wide consumption. The devices clustered include air conditioners, electric vehicle chargers, dishwashers, garbage disposals, dryers, freezers, fridges, furnaces, garage lights, microwaves, ovens, and stove ranges. These devices were selected due to both the abundance of data for many of them as well as some of the interesting properties featured with each. For in-

stance, electric vehicle chargers represent a relatively new household commodity and the researchers thought they may show more diverse load-signatures due to manufacturing differences than other appliance types. Furthermore, many of the selected appliances are relatively behavior independent meaning that their usage patterns may correlate less to occupant behavior than some other appliance types. The major exceptions to this are the oven, stove range, and microwave. Appliances such as dryers, dishwashers, garage lights, and disposals require occupant interaction to be turned "ON" or "OFF" but likely have less direct behavioral differences then cooking equipment which is manually set at different temperatures befitting cuisine.

Following the algorithm outlined in section 4.1, hyperparameters were set as: $t = 720, 1440, 4320$, $c = 2$, and $k = 20$. The three values for t denote that the process was repeated for three separate experimental scenarios. Clustering on 720 minutes, 1440 minutes, and 4320 minutes respectively. For the preprocessing step, each appliance timeseries was downsampled inclusively (seconds $1 \rightarrow 60$ were included in each minute as opposed to $1 \rightarrow 59$) for 1 minute long periods with each new entry representing the average of the minute. After removing "OFF" appliances, the remaining samples were scaled using a standard scalar as recommended in Section 4.2. The clustering process loop was then executed for each appliance and the total aggregate load.

4.2 Outlier Detection Results

The resulting outlier appliances are shown in Table 2. In the table, the outliers for each appliance are shown to the right of the corresponding appliance name. The columns labeled numerically indicate the number of minutes that the clustering process was run on as outlined in Section 4.1.

The results of the outlier identification notably vary for some appliances during different time lengths. This may at first seem unusual but is generally explained by more appliances qualifying for the clustering algorithm by being "ON" during the period and by appliances showing larger patterns of overarching behavior over longer windows.

The introduction of new samples into the clustering pool with the increasing time length deserves further consideration. When an appliance that is not present in a short time length clustering trial is present in a longer period

$t(minutes)$:	720	1440	4320
Air Conditioner	House 3039	House 3039	House 3039
Car Charger	House 9053	House 9053, House 3000*	House 9053, House 3000*
Dishwasher	House 9019	House 1417	House 4031
Garbage Disposal	House 3039	House 9922, House 3456	House 5587
Dryer	House 9278	House 3996, House 9019	House 9053
Freezer	House 1240	House 3000	House 142
Fridge	House 5982	House 3700	House 4031
Furnace	House 1240	House 8565	House 1240, House 5746
Garage	House 6139	House 5997	House 27
Household	House 2096	House 7951	House 1417
Heater	House 1240*	House 3700, House 5982	House 3700
Microwave	House 1642	House 1642	House 661
Oven	No outlier	House 9922	House 3456
Stove Range	House 5587	House 1222	House 1222

Table 2: The outliers identified for each appliance type.
* The outlier identified was the second such appliance in the same house.

it is worth pointing out that this means that the appliance is likely "OFF" for a longer duration of the clustering window than the other samples being clustered. This may artificially lead it to appear an outlier even when scaled. Its behavior is expected to look different than other samples which were "ON" for longer portions of the clustering window. While this is a valid concern, there is a similar level of uncertainty in the smallest time-length clustering. Without inverting the check to exclude any appliance that is *ever* "OFF" during the window (something that would prohibit the clustering of all but constant use appliances like refrigerators), there is going to be variance in the portion of the time where each sample is "ON".

Another noteworthy result of the clustering experimentation is the detection of multiple outliers for a specific time length. In Table 2, this is shown as two homes in the same column. This phenomenon occurs when a cluster that had two samples within it, is separated further into two clusters with only one sample each. For example, if the $t = 1440$, $c = 3$ iteration of the algorithm identified a cluster with the dryer from House 3996 and the dryer from House 9019, the $c = 4$ iteration could split this two-sample cluster into two individual clusters with one dryer each.

4.3 Sample Comparison

After identifying the various device type outliers, it is possible to visually affirm the functionality of the algorithm. By plotting outliers against random subsets of their device class (all unscaled), outliers can be visually confirmed. For example: Figure 2 and Figure 3. Figure 2 shows a comparison of the identified outlier fridge–that of house 3700–and a random subset of the other fridges present in the dataset. As can be seen, the outlier, plotted in the dotted black line, shows atypical behavior as compared to the other samples. A fridge is an example of an inductive load device, and the identified outlier shows a much "flatter" load signature than the other samples in its category. It is likely still identifiable as a fridge, due to their unique load pattern, but it is clearly an outlier compared to the other samples.

A similar analysis is made in Figure 3 which shows a comparison of the outlier furnace, again in the dotted black line, versus a random subset of the other furnaces in the dataset. Again, the identified outlier is clearly anomalous as compared to the other furnace samples in the dataset. A furnace is a type of resistive load and is expected to show a "step" behavior as discussed in [44]. Again, the identified outlier is far "flatter" than the load signatures of the other appliance samples.

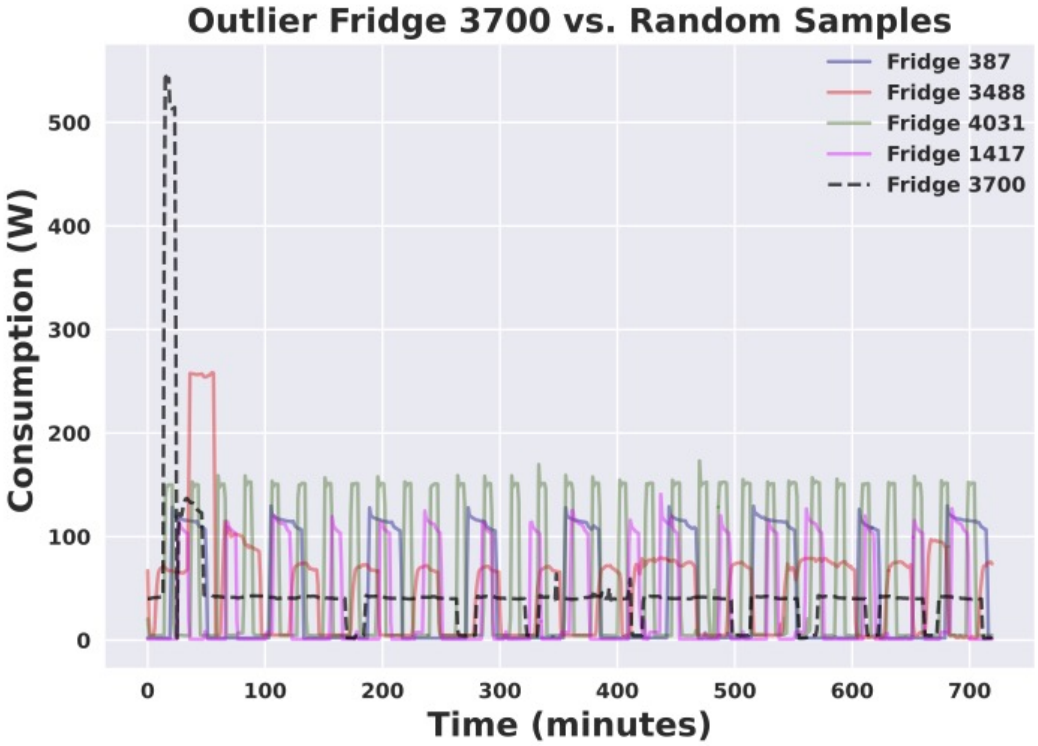


Figure 3: A comparison of the identified fridge outlier (black) versus a random subset of other fridges.

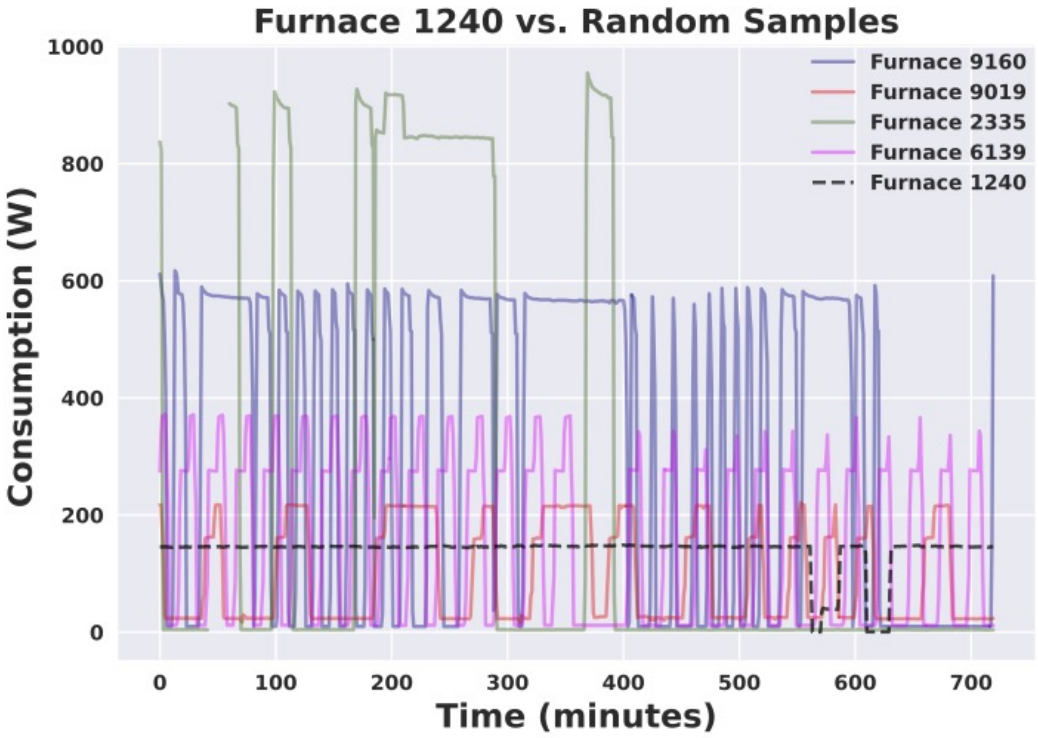


Figure 4: A comparison of the identified furnace outlier (black) versus a random subset of other furnaces.

A final evaluative step was taken to verify that the algorithm remains successful across time lengths. As noted in Table 2 and its related discussion in Section 4.2, it was not uncommon for the algorithm to identify different outlier samples for the same appliance when clustering on different time lengths. In order for the outlier detection algorithm to remain useful, it is important that when comparing two different outliers of the same appliance type both *are* outliers. While it would be convenient if the same outlier was always identified, what is important is that the identified samples are always anomalous. Figure 4 shows a second anomalous fridge, again in the dotted black line, compared to a random subset of other fridges in the dataset. While the second fridge clearly shows a different behavior pattern than the first outlier fridge shown in Figure 2 (a far more rapid load pattern), it is still clearly an outlier for its device category.

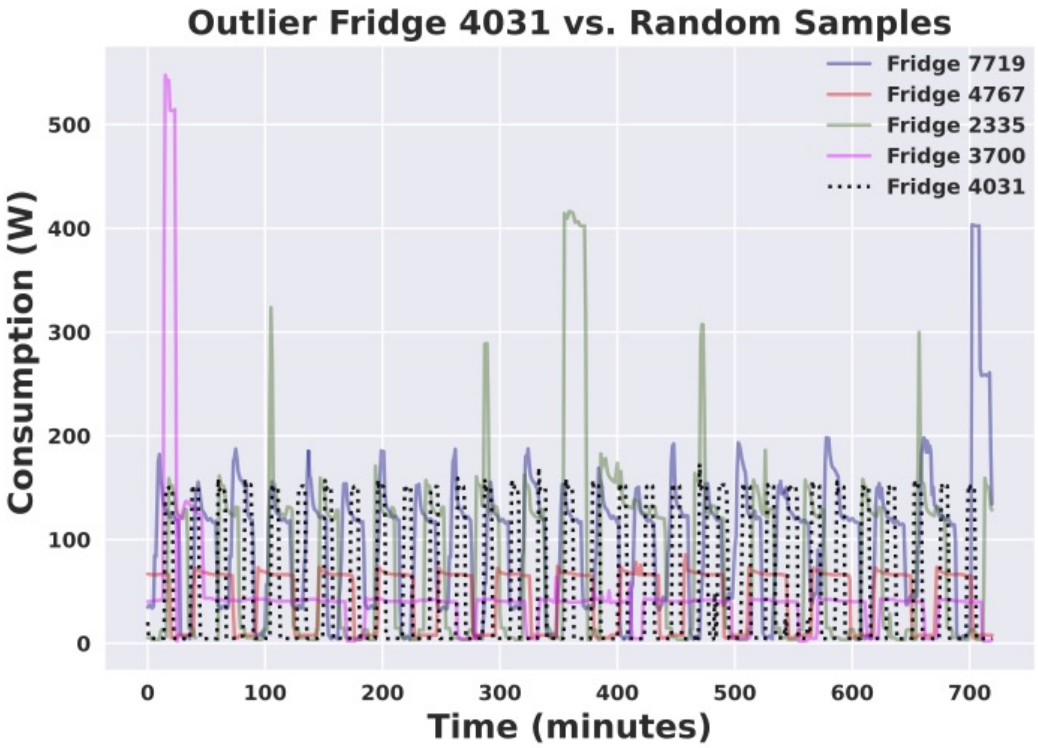


Figure 5: A comparison of another outlier fridge. Outlier (black) vs. random subset of other fridges.

5 Conclusion

5.1 Contributions

The algorithm proposed in this article is significant for its demonstration of the ability of common, even simple unsupervised clustering techniques to effectively identify outliers in NILM datasets. This research was motivated by an interest in determining the representativeness of commonly used NILM datasets. As discussed, NILM datasets have a wide variety of applications which necessitate that they contain representative real-world appliance data. As an avenue to determine the representativeness of the data within the NILM datasets, an outlier detection algorithm was constructed. The algorithm was constructed from three main steps namely: soft-Dynamic Time Warping, Timeseries K-Means Clustering, and Barycenter Averaging. In order to evaluate the efficacy of the algorithm, outliers were detected for 13 different appliance types using a 50 home subset of the Dataport NILM dataset. As discussed in the previous section, the outliers identified by the algorithm were compared to random subsets of the dataset used. This evaluation demonstrated that the algorithm effectively identified atypical appliances.

Using the developed algorithm, there are numerous avenues for future research and applications as discussed in the following subsection.

5.2 Future Work

Quantifying Outlier Impact Perhaps the most obvious future line of inquiry based on this algorithm is determining the extent to which the outliers in NILM datasets matter. While the algorithm is clearly able to identify anomalous appliances, it is not clear to what extent these atypical samples impact NILM algorithms' generalization capabilities. In order to quantify this impact, future research could take an existing NILM algorithm–trained on a publicly available NILM dataset–record the reported statistics for the algorithm when the outliers remain in the dataset, and then remove the outliers before rerunning the original experimentation to get new numeric values. For instance, if a NILM publication reports a 92% accuracy for disaggregating appliances, and states it was trained on the SMART* dataset, it would be fruitful to remove the

outliers from SMART* and retrain the same algorithm in order to gauge whether its effectiveness increases or decreases as a result of removing the outliers.

Demand-side Management A second application of the algorithm is its deployment by demand-side management companies in order to provide real-time similarity metrics for consumer appliances. Given the algorithm's current state, it is able to effectively identify several types of anomalous appliance behavior. These anomalous behavior patterns can be used with existing NILM techniques to allow demand-side management to offer real-time behavior analysis of consumer appliances. This could take the form of energy providers being able to inform customers when their appliances' behavior appears more similar to that of an outlier device than a typical load.

Outlier Cause Analysis A third application of the outlier detection algorithm is the identification of underlying outlier causes. Effectively research into the causes of anomalous appliance behavior could be conducted through identifying outliers and then comparing the outliers for similar traits or faults. This application may elucidate further opportunities for hardware and manufacturing improvements as well as help customers make informed purchases.

Beyond lines of inquiry that focus on the application of the algorithm as-is, there are also promising opportunities for the improvement of the algorithm itself. Some of these opportunities include:

Definitions of Outlier Behavior Due to the varied nature of consumer appliances, it is not obvious that formal mathematical definitions of an outlier would facilitate the successful identification of atypical devices. That being said, as was shown, the current algorithm lacks a specific definition for outlier behavior and accordingly can identify outliers expressing different behavior. By implementing a tested mathematical definition of outlier behavior or through the construction of a semantic definition of an outlier, it would be possible to direct the algorithm towards more consistent identification of specific types of atypical devices.

Comparison of Clustering Techniques The algorithm is currently implemented using Timeseries K-Means Clustering, however, this is far from the only clustering approach. To this point, there is again a great opportunity for future research which can analyze the appropriateness of different clustering techniques for the process of identifying outliers. While easily interpreted and conducive to centroid analysis, Timeseries K-Means does strictly partition clusters into a specified number of groups in a way that does not allow for an appliance to be similar to multiple groups. Other clustering approaches such as agglomerative clustering lack this rigidity and may offer a better intuition into the overall behavior categorization of different appliance samples. Furthermore, while it was previously mentioned that non-clustering approaches of outlier detection are suspected to face challenges in the definition of global hyperparameters for all appliance types, these approaches may still offer a more computationally efficient basis for outlier detection.

Beyond applications to future research, this paper outlines a novel approach to the identification of outlier appliances in energy datasets. The purpose of this contribution is to further improve energy conservation techniques as well as offer an assistive hand to other researchers with like-minded queries.

References

[1] B. Neenan, J. Robinson, and B. R., "Residential electricity use feedback: A research synthesis and economic framework," Electric Power Research Institute, 2009.

[2] W. Kong, Z. Dong, B. Wang, J. Zhao, and J. Huang, "A practical solution for non-intrusive type ii load monitoring based on deep learning and post-processing," *IEEE Transactions on Smart Grid*, vol. PP, pp. 1–1, May 2019. DOI: 10.1109/TSG.2019.2918330.

[3] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "Nilmtk," *Proceedings of the 5th international conference on Future energy systems*, Jun. 2014. DOI: 10.1145/2602044.2602051. [Online]. Available: http://dx.doi.org/10.1145/2602044.2602051.

[4] G. W. Hart, "Prototype nonintrusive appliance load monitor," MIT Energy Laboratory, Tech. Rep. Progress Report 2, Sep. 1985.

[5] S. Singh and A. Yassine, "Big data mining of energy time series for behavioral analytics and energy consumption forecasting," *Energies*, vol. 11, no. 2, 2018. [Online]. Available: https://www.mdpi.com/1996-1073/11/2/452.

[6] K. Yan, X. Wang, Y. Du, N. Jin, H. Huang, and H. Zhou, "Multi-step short-term power consumption forecasting with a hybrid deep learning strategy," *Energies*, vol. 11, no. 11, 2018.

[7] "Energy demand side management within micro-grid networks enhanced by blockchain," *Applied Energy*, vol. 228, pp. 1385–1398, 2018, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2018.07.012.

[8] O. Tan, J. Gómez-Vilardebó, and D. Gündüz, "Privacy-cost trade-offs in demand-side management with storage," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1458–1469, 2017. DOI: 10.1109/TIFS.2017.2656469.

[9] W. Kong, Y. Xu, Z. Y. Dong, D. J. Hill, J. Ma, and C. Lu, "An extended prototypical smart meter architecture for demand side management," in *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, 2015, pp. 1008–1013. DOI: 10.1109/INDIN.2015.7281873.

[10] A. Yassine, S. Singh, and A. Alamri, "Mining human activity patterns from smart home big data for health care applications," *IEEE Access*, vol. 5, pp. 13 131–13 141, 2017. DOI: 10.1109/ACCESS.2017.2719921.

[11] D. Yan, W. O'Brien, T. Hong, X. Feng, H. Burak Gunay, F. Tahmasebi, and A. Mahdavi, "Occupant behavior modeling for building performance simulation: Current state and future challenges," *Energy and Buildings*, vol. 107, pp. 264–278, 2015, ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2015.08.032. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778815302164.

[12] W. Kleiminger, C. Beckel, T. Staake, and S. Santini, "Occupancy detection from electricity consumption data," ser. BuildSys'13, Roma, Italy: Association for Computing Machinery, 2013, pp. 1–8, ISBN: 9781450324311. DOI: 10.1145/2528282.2528295. [Online]. Available: https://doi.org/10.1145/2528282.2528295.

[13] J. M. Alcalá, J. Ureña, Á. Hernández, and D. Gualda, "Assessing human activity in elderly people using non-intrusive load monitoring," *Sensors*, vol. 17, no. 2, 2017.

[14] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," *Proc. SustKDD.*, Jan. 2012.

[15] *What makes a product energy star?* [Online]. Available: https://www.energystar.gov/products/what_makes_product_energy_star.

[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, 2011, pp. 59–62.

[17] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, pp. 1–20, Nov. 2019. DOI: 10.1371/journal.pone.0224365. [Online]. Available: https://doi.org/10.1371/journal.pone.0224365.

[18] D. Yan, W. O'Brien, T. Hong, X. Feng, H. Burak Gunay, F. Tahmasebi, and A. Mahdavi, "Occupant behavior modeling for building performance simulation: Current state and future challenges," *Energy and Buildings*, vol. 107, pp. 264–278, 2015, ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2015.08.032. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778815302164.

[19] J. Kelly and W. Knottenbelt, "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*, ser. BuildSys '15, Seoul, South Korea: Association for Computing Machinery, 2015, pp. 55–64, ISBN: 9781450339810.

[20] M. J. Johnson and A. S. Willsky, *Bayesian nonparametric hidden semi-markov models*, 2012. arXiv: 1203.1365 [stat.ME].

[21] W. Kleiminger, C. Beckel, T. Staake, and S. Santini, "Occupancy detection from electricity consumption data," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, ser. BuildSys'13, Roma, Italy: Association for Computing Machinery, 2013, pp. 1–8, ISBN: 9781450324311. DOI: 10.1145/2528282.2528295. [Online]. Available: https://doi.org/10.1145/2528282.2528295.

[22] D. Chen, S. Barker, A. Subbaswamy, D. Irwin, and P. Shenoy, "Non-intrusive occupancy monitoring using smart meters," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, ser. BuildSys'13, Roma, Italy: Association for Computing Machinery, 2013, pp. 1–8, ISBN: 9781450324311. DOI: 10.1145/2528282.2528294. [Online]. Available: https://doi.org/10.1145/2528282.2528294.

[23] V. Kekatos, G. Wang, A. J. Conejo, and G. B. Giannakis, "Stochastic reactive power management in microgrids with renewables," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3386–3395, 2015. DOI: 10.1109/TPWRS.2014.2369452.

[24] J. Kelly and W. Knottenbelt, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes," *Scientific Data*, vol. 2, no. 150007, DOI: 10.1038/sdata.2015.7.

[25] S. Noor, W. Yang, M. Guo, K. H. van Dam, and X. Wang, "Energy demand side management within micro-grid networks enhanced by blockchain," *Applied Energy*, vol. 228, pp. 1385–1398, 2018, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2018.07.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261918310390.

[26] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for non-intrusive load monitoring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[27] A. Yassine, S. Singh, and A. Alamri, "Mining human activity patterns from smart home big data for health care applications," *IEEE Access*, vol. 5, pp. 13 131–13 141, 2017. DOI: 10.1109/ACCESS.2017.2719921.

[28] M. Wurm and V. Coroama, "Poster abstract: Grid-level short-term load forecasting based on disaggregated smart meter data," *Computer Science - Research and Development*, vol. 33, pp. 1–2, Sep. 2017. DOI: 10.1007/s00450-017-0374-3.

[29] T. Wijaya, S. Humeau, M. Vasirani, and K. Aberer, "Residential electricity load forecasting: Evaluation of individual and aggregate forecasts.," School of Computer and Communication Sciences, Tech. Rep., 2014.

[30] C. Dinesh, S. Makonin, and I. V. Bajić, "Residential power forecasting using load identification and graph spectral clustering," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 11, pp. 1900–1904, 2019. DOI: 10.1109/TCSII.2019.2891704.

[31] Y. Hong, Y. Zhou, Q. Li, W. Xu, and X. Zheng, "A deep learning method for short-term residential load forecasting in smart grid," *IEEE Access*, vol. 8, pp. 55 785–55 797, 2020. DOI: 10.1109/ACCESS.2020.2981817.

[32] S. Desai, R. Alhadad, A. Mahmood, N. Chilamkurti, and S. Rho, "Multi-state energy classifier to evaluate the performance of the nilm algorithm," *Sensors*, vol. 19, p. 5236, Nov. 2019. DOI: 10.3390/s19235236.

[33] K. S. Barsim, R. Streubel, and B. Yang, "An approach for unsupervised non-intrusive load monitoring of residential appliances," Jun. 2014.

[34] Y.-H. Lin, M.-S. Tsai, and C.-S. Chen, "Applications of fuzzy classification with fuzzy c-means clustering and optimization strategies for load identification in nilm systems," in *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, 2011, pp. 859–866. DOI: 10.1109/FUZZY.2011.6007393.

[35] M. Aiad and P. H. Lee, "Energy disaggregation of overlapping home appliances consumptions using a cluster splitting approach," *Sustainable Cities and Society*, vol. 43, pp. 487–494, 2018, ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2018.08.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670718301197.

[36] F. Jazizadeh, B. Becerik-Gerber, M. Berges, and L. Soibelman, "Unsupervised clustering of residential electricity consumption measurements for facilitated user-centric non-intrusive load monitoring," Jun. 2014, pp. 1869–1876. DOI: 10.1061/9780784413616.232.

[37] F. Jazizadeh, B. Becerik-Gerber, M. Berges, and L. Soibelman, "An unsupervised hierarchical clustering based heuristic algorithm for facilitated training of electricity consumption disaggregation systems," *Advanced Engineering Informatics*, vol. 28, Oct. 2014. DOI: 10.1016/j.aei.2014.09.004.

[38] H. Liu, Q. Zou, and Z. Zhang, "Energy disaggregation of appliances consumptions using ham approach," *IEEE Access*, vol. 7, pp. 185 977–185 990, 2019. DOI: 10.1109/ACCESS.2019.2960465.

[39] C. Puente, R. Palacios, Y. González-Arechavala, and E. F. Sánchez-Úbeda, "Non-intrusive load monitoring (nilm) for energy disaggregation using soft computing techniques," *Energies*, vol. 13, no. 12, 2020, ISSN: 1996-1073. [Online]. Available: https://www.mdpi.com/1996-1073/13/12/3117.

[40] C. Dinesh, S. Makonin, and I. V. Bajić, "Residential power forecasting using load identification and graph spectral clustering," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 11, pp. 1900–1904, 2019. DOI: 10.1109/TCSII.2019.2891704.

[41] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[42] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series.," in *KDD workshop*, Seattle, WA, USA: vol. 10, 1994, pp. 359–370.

[43] M. Cuturi and M. Blondel, *Soft-dtw: A differentiable loss function for time-series*, 2017. DOI: 10.48550/ARXIV.1703.01541. [Online]. Available: https://arxiv.org/abs/1703.01541.

[44] S. Barker, S. Kalra, D. Irwin, and P. Shenoy, "Empirical characterization and modeling of electrical loads in smart homes," in *2013 International Green Computing Conference Proceedings*, 2013, pp. 1–10. DOI: 10.1109/IGCC.2013.6604512.