

A Snapshot of NILM: Techniques and Tests of Non-Intrusive Load Monitoring for Load Shape Development

Justin Elszasz, Navigant Consulting Inc., Washington, DC

Tracy Dyke-Redmond, Eversource, Westwood, MA

Justin Spencer, Kathleen Ward, Daniel Zafar, Ken Seiden, Terese Decker,

Chris Newton, Navigant Consulting Inc., Boulder, CO

ABSTRACT

For utilities designing and implementing peak demand reduction programs, it is important to have interval data to understand not only how much energy is being used, but when it is being used, and ideally, for what equipment. As part of a residential baseline study in the Northeast, a group of evaluators, energy efficiency program administrators, and other stakeholders tested the validity of some non-intrusive load monitoring (NILM) technologies. The study tested optical sensors that “watch” the utility meter, digitize the energy usage data, and transmit it to software, which in turn analyzes the data and develops disaggregated load shapes for household equipment. The study tested three different software approaches: a proprietary algorithm, an open-source algorithm (Makonin 2015), and an econometric model (Parti 1980). This study considers the accuracy of the disaggregated data for different types of equipment. While disaggregation was successful for some end-uses (generally ones correlated strongly with weather), many end uses could not be disaggregated in households whose data had not been used for training the algorithms. The results of this study will be widely applicable to evaluators seeking to use cost effective approaches for gathering interval energy use data and developing load shapes, particularly for the residential sector but potentially also for small businesses.

Introduction

Understanding residential energy end-use behavior is critical for energy efficiency program design and evaluation. The electric grid is becoming increasingly dynamic, where at any given moment, varying amounts of solar photovoltaic, solar thermal, wind, or load “shedding” events may be combined with coal, natural gas, and nuclear energy to serve the instantaneous demand, all of which have costs and capacity that vary in time. Demand response events, themselves a resource, are often the most cost effective way for a utility to shed load and reliably meet customer demand. But knowing when and what types of demand-response events to trigger (i.e., which appliances to incentivize customers to turn off, and for how long) and how much utilities should pay customers to reduce their demand is a complex socio-techno-economic question that depends on regional demographics, housing characteristics, and energy prices.

To design cost-effective and robust demand response programs, utilities must understand end-use behavior in more detail than ever before. If a utility has an accurate estimate of what fraction of total load is from various appliances and at a particular instant, it can predict how much instantaneous demand could potentially be shed and total energy savings that could be achieved by triggering an event, thereby ensuring power reliability without purchasing more supply.

Program evaluators equally benefit from understanding end-use behavior at a more granular level. Energy savings and demand reductions for particular measures could be obtained directly through load disaggregation without reliance on program-wide assumptions. Alternatively, load disaggregation could be used to refine those assumptions. For example, evaluation typically uses constant assumptions

of coincidence factors across the program, meaning it assumes that the interplay of various appliances with heating and cooling loads is constant over large periods of time (often separate summer and winter periods) and for all instances of the same building and space type combination. Armed with hourly or sub-hourly load profiles of appliances including the heating and cooling loads, evaluators could offer more accurate measurements of the effects of efficiency programs, both in terms of total kilowatt-hour savings and power (kilowatt) demand reductions.

Load Disaggregation or Non-Invasive Load Monitoring

The most accurate means for measuring hourly or sub-hourly load profiles for end-uses is still submetering – using current transducers or other logger types to independently measure the power draw of appliances over time. The measurement accuracy is limited only by the sensor device itself and the frequency at which the device records measurements for analysis. Submetering has two related key drawbacks: 1) cost and 2) intrusiveness. Costs can quickly escalate when performing highly-accurate appliance load monitoring across multiple houses. The evaluation team considers the benefits and burdens of submetering vs. NILM in “Duckhunt! Benefits and risks of load disaggregation and end use metering for determining end use loadshapes” (Decker 2017). In a submetering approach, each house circuit and/or individual appliance must have a dedicated sensor, all of which can feed into a central logging gateway for communication over the internet, or be stored in an onsite logger. While new logging and sensing technology has made end use sub-metering much cheaper than it was ten years ago, it still costs between \$50 and \$500 per metered device. Depending on the size of the house and number of appliances or electronics desired to be measured, the cost of hardware can easily exceed a thousand dollars for one house. Beyond just the hardware, the submetering process is inherently intrusive. The evaluation team, along with an electrician, must enter the home and have access to the electrical panel, any sub-panels, and each appliance, light, or electronic device being studied, thereby requiring that the homeowner be home for installation, any maintenance or troubleshooting, and for removal of the hardware. Without an incentive to participate, this type of study is overly burdensome to conduct from the customer’s perspective.

An exciting prospect for load shape development is the use of computer algorithms to detect individual appliance loads from an aggregate whole house data. Load disaggregation, or non-intrusive load monitoring (NILM), requires only the time series of the whole house energy/power draw and its characteristics to estimate individual appliance loads. NILM algorithms use various strategies to detect appliance states and the accompanying power draw of each appliance. In general, due to the reliance on machine learning techniques, the algorithm must first be “trained;” that is, the algorithm must first use a dataset that includes the ground-truth about what appliances are on and off and any given instant to learn what each appliance’s load signature looks like in the whole-house readings. NILM must overcome this hurdle to provide accurate disaggregation for households on which the algorithm has not been previously trained. This paper considers whether or not a subset of current NILM technologies could overcome that hurdle.

The use of advanced metering infrastructure (AMI), or smart meters, already being installed by utilities to provide this data stream is particularly appealing for its simplicity (new, dedicated hardware not needed). But the current state-of-the-art algorithms require much more frequent measurements to disaggregate end uses than the hourly or 15-minute interval data smart meters currently provide; an hourly reading of a house’s kilowatt-hour usage in that hour is insufficient for detecting what portion of that draw went to individual loads. It is generally understood that higher frequency data will allow for the disaggregation of more end uses and end uses with relatively smaller power draws. While advanced AMI may provide sufficient data for disaggregation in the future, AMI data were not available for this study at any resolution.

Approach to NILM

This evaluation of NILM technology was part of a broader baseline study in which the Massachusetts Program Administrators are developing new hourly load profiles for a range of energy end-uses to inform future program design and evaluation. NILM was selected to reduce measurement costs, and the team took a phased approach to the baseline study, with the first phase intended to determine if using NILM was feasible for load disaggregation. The evaluators did not trust NILM to produce accurate results without first testing the approach, even for what might be considered the current best-in-class NILM devices sampling at a frequency of more than once per second. The original study designed assumed that a large sample of NILM sites could be bias corrected using a nested sample of end use submetering sites. This would allow a much wider sample to be included in the study, incorporating a broad range of customer demographics and usage behaviors and reducing the likelihood of sample bias. This approach depended on finding a low cost NILM hardware/software combination that could deliver reasonable accuracy. If disaggregation proved to be successful in the first phase, the team anticipated using NILM in subsequent phases in order to reduce the number of sites using more costly but highly accurate sub-metering approach.

Hardware & Data

During the pilot phase of the baseline study for developing appliance loadshapes, the team used two sets of hardware: one device for measuring the whole-house energy use (to be used in conjunction with disaggregation algorithms) and one sub-metering device for measuring the actual usage of each appliance, or “ground truth,” to serve as a basis for comparison regarding metric accuracy. As stated previously, AMI data were not available for this study. As described earlier, the appeal of NILM technology is to reduce the amount of hardware needed and simplify the installation process when measuring end-use loadshapes. The simpler, less expensive hardware scheme could mean a larger sample size for the same overall budget, thereby improving the representativeness and (potentially) the overall accuracy of the baseline study after the pilot phase. Towards that end, our team selected an optical sensor that attaches to the outside of a utility meter and “reads” the usage at higher frequency than typical AMI data. It then transmits the data via the household’s wireless network.



Figure 1: Home energy meters with attached sensor. The sensor calculates energy usage using patented optical technology applied to the spinning wheel.

The team installed whole-house meter optical sensors and intensive submetering in 23 houses in Massachusetts for the purposes of training disaggregation algorithms and for measuring the accuracy of each disaggregated appliance load shape. The whole-house optical-sensors collected readings approximately every 30 seconds while the submetering hardware (current transducer-based) collected readings every minute. The whole-house 30-second interval data was therefore downsampled to 1-minute to match the time resolution of the submetering data.

Data collection began in late August 2016 for some households and by mid-September 2016 data were being collected at all 23 sites. The pilot phase, during which NILM was being evaluated for potential use in the second phase of the baseline study, concluded at the end of November 2016.

Methodologies Tested

To disaggregate the whole-house energy use measurements into end-use load profiles, the evaluation team compared the accuracy of three different disaggregation approaches. First, the team considered a proprietary algorithm from a third-party NILM software vendor with expertise in load disaggregation. The team also used an open-source algorithm to attempt disaggregation. Finally, the team implemented an econometric model to derive hourly load profiles as a % of total load of a group of houses.

To provide a fair and neutral assessment of accuracy of the proprietary algorithm, the open-source algorithm, and the econometric model, data for six sites (hereafter “test sites”) were set aside and unused until the team conducted final accuracy evaluations. This prevented use of the same data for both algorithm development and accuracy measurements, which could result in undeservedly high accuracy estimates.

Proprietary Algorithm

There are several companies that specialize in using software to disaggregate whole-house electricity data into appliance loadshapes. Some companies perform disaggregation for consumers and building owners by providing web-based dashboards that estimate the energy use of appliances for homeowners. These software solutions pair with current transducer-based home energy monitors as the source of their data. Others are attempting disaggregation using utility smart meter data (typically 15-minute interval) to provide appliance-level insights to utilities for demand-side management programs, measurement and verification, or load forecasting.

Choosing a NILM software vendor and getting them on board was not easy. The published data on NILM provider software accuracy was sparse or non-existent. Perhaps tellingly, multiple providers ultimately pulled out of participation in this study, generally because this was not a use case that they were interested in. The team ultimately found one proven third-party NILM software vendor that specializes in disaggregating 15-minute interval meter data for utilities that was willing to participate in the study. To perform the evaluation, the vendor was provided with only 1-minute interval whole-house demand data. Initially, the vendor was not provided with any submetering data but subsequently the vendor received some submetering data to retrain their algorithm and potentially increase their accuracy.

Open-Source Algorithm

The team leveraged an open-source disaggregation algorithm called SparseNILM developed by Stephen Makonin (Makonin 2015). The code is available on the code-hosting website Github for public

use.¹ The algorithm at its core is a Hidden Markov Model (HMM), a model that builds in sequential dependency by estimating the probability of transitioning from one state to the next, in this case, from a house transitioning from having one set of appliances turned on to another state in which a different set of appliances is on. Another set of probabilities determines the likelihood that, given a household state, the whole-house current draw is the sum of the currents of the appliances that are not off. Through these sets of probabilities, the state of appliances in the household can be deduced given only the whole-house current draw.

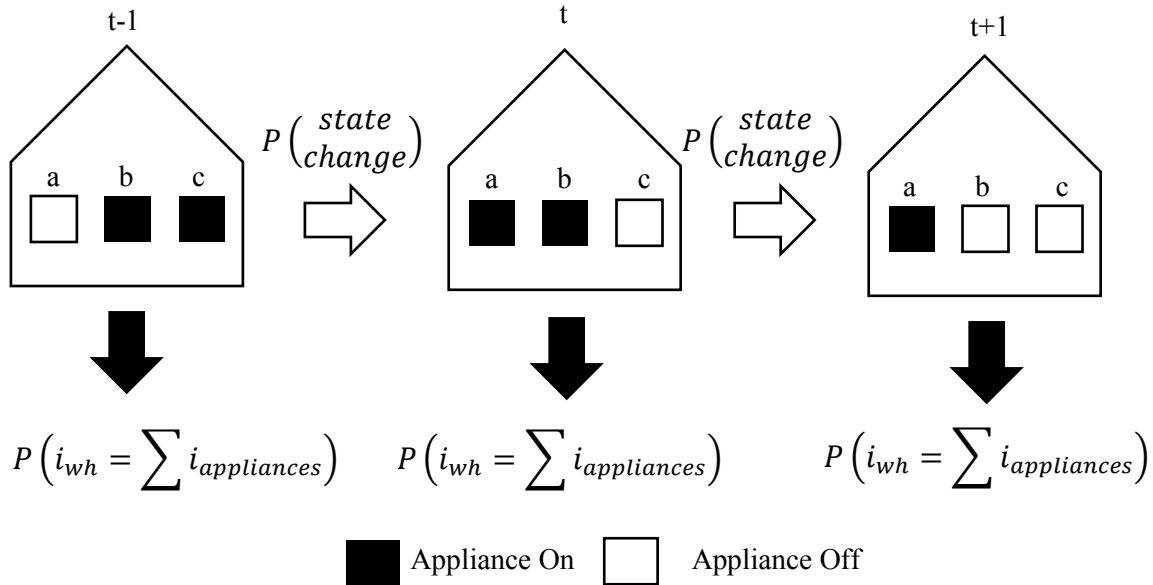


Figure 2. Hidden Markov Model for load disaggregation. The observed values for the model are the time series of whole-house current draw. The state of appliances in the house at each time step is inferred through two sets of probabilities – the probability that the whole-house current draw is comprised of current draws for a particular set of appliances that are in an on-state and the probability of a particular set of appliances being on given the set of appliances that were on at the previous time step.

Econometric Model

The econometric model used for our disaggregation was a form of Conditional Demand Analysis (CDA), an approach originated by Michael Parti and Cynthia Parti in 1980. CDA uses residential whole-house data and variation in appliance ownership across households to identify end use loads. The theory behind the conditional demand model is that a household’s total electric usage is the sum of its end uses. By exploiting variation in end uses across a group of households, the econometric model can isolate individual end use loads. Traditionally, CDA employs electric billing data to disaggregate loads at the monthly level. However, the same methodology can be used on interval data to disaggregate loads at the hourly level.

Formally, the hourly conditional demand model is specified as follows (Blaney 1994):

$$y_{it} = \sum_{j=1}^J \beta_{ijt} \cdot D_{ijt} + \sum_{k=1}^K \lambda_{ikt} \cdot X_{ikt} + e_{it}$$

Where:

¹ For further details of algorithm and code, see <https://github.com/smakonin/SparseNILM>

y_{it} = the electricity use of household i during period t.

β_{ijt} = the estimated electricity demand of appliance j in household i during period t

D_{ijt} = a dummy variable taking the value 1 if household i owns appliance j during period t, and 0 otherwise

λ_{ikt} = a vector of estimated effects of K demographic and weather variables on household demand, specific to household i during period t

X_{ikt} = a vector of K demographic and weather variables, specific to household i during period t

e_{it} = the error term for household i during period t

Estimating this whole-house regression using end-use dummy variables results in estimates of parameters $\hat{\beta}$ defining the expected value of any end-use load conditional on the observables X, with an associated covariance matrix $Cov(\hat{\beta})$.

Given that individual appliance metering data was available, the team took the conditional demand model a step further and applied a form of Bayesian updating, similar to the approach outlined by Caves (1987). With end-use observations, the team applied the exact same end-use model used in the whole-house CDA specification with a slight modification. In this case, the end-use observation was the dependent variable, thereby recovering another set of estimates of the parameters defining the expected value of any end-use load conditional on the observables X. This model is specified as follows:

$$z_{ijt} = \sum_{k=1}^K \lambda_{ijk} \cdot X_{ikt} + e_{ijt}$$

Where:

z_{ijt} = the electricity use of appliance j in household i during period t.

λ_{ikt} = a vector of estimated effects of K demographic and weather variables on household demand, specific to household i during period t

X_{ikt} = a vector of K demographic and weather variables, specific to household i during period t

e_{ijt} = the error term for appliance j in household i during period t

We combined our estimates into a weighted mean with greater weight placed on the estimates with smaller variance. This follows Becker and Wu (2007), and is the form of Bayesian updating advocated in Caves (1987).

Evaluation Metric

In this study, the team recognized the potential to “true-up” disaggregated loadshapes to the proper magnitude with ratios. This approach partially mitigates the potential that the algorithm might inaccurately capture the total quantity of energy use, even though it may have produced usable loadshapes for specific equipment types. This approach does a good job of correcting for bias in a NILM result, e.g. if the NILM results were consistently too high or too low for a given end use.

The hours of the pilot phase were divided into bins by weekend/weekday and period of the day (morning, midday, afternoon, evening, and night). The ratios for truing up the appliance load are calculated for each of the bins as follows:

$$r_{bin} = \frac{\bar{y}_{bin, logged}}{\bar{y}_{bin, disagg}}$$

Where $\bar{y}_{bin, logged}$ is the mean logged (i.e. “ground truth”) appliance demand for the bin and $\bar{y}_{bin, disagg}$ is the mean of the appliance demand for the bin estimated by the disaggregation algorithm. The true-up ratio for the bin is then applied to the predicted appliance demand from the algorithm.

Using this corrected appliance load estimate, the coefficient of variation (CV), which in this case is the root mean square error (RMSE) divided by the mean logged demand, is calculated for each bin. .

$$RMSE_{bin, disagg} = \sqrt{\frac{\sum_{i=1}^n (r_{bin} \hat{y}_i - y_i)^2}{n-1}}$$

$$CV_{bin, disagg} = \frac{RMSE_{bin, disagg}}{\bar{y}_{bin, logged}}$$

Typically, the CV is calculated around a mean value, where the squared differences between each measurement and the mean are used.

$$CV_{bin, mean} = \frac{RMSE_{bin, logged}}{\bar{y}_{bin, logged}} = \frac{\sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y}_{bin, logged})^2}{n-1}}}{\bar{y}_{bin, logged}}$$

In this study, the disaggregation CV’s are compared to the CV around the means. If the disaggregation CV is lower than the CV around the mean, then the disaggregation result is judged to have provided a better estimate of the appliance load than simply using the mean itself. This is equivalent to an evaluator trying to decide whether to use a ratio estimator approach or a simple mean to calculate a realization rate for a program. If the ratio estimator has a better CV, then the ratio estimator provides a better estimate and the prior estimates of savings provide useful statistical power that can be used to reduce the sample size. In this case, if the ratio estimator CV is better than the straight mean CV, then the NILM algorithm offers a better prediction of the energy use of an appliance compared to the straight mean.

Results

In terms of whole-house power measurement, the tradeoffs of the particular whole-house device our team used were the relatively low frequency readings (compared to the sampling rates used in research-grade NILM development) and lack of additional power characteristics (e.g. power factor, frequency, etc.) The selected NILM hardware was capable of approximately 30-second interval data that consisted only of the instantaneous power draw in Watts (though, as noted, this was downsampled further). A further limitation of the NILM hardware was a measurement “floor.” Because the device sensed the power draw displayed by the utility meter at regular intervals, there was a power draw threshold below which the device could not determine the power draw and would therefore report that minimum threshold draw for the time interval. While these aforementioned limitations were known features of the device, the team also encountered poor data quality, including intermittency and spikes. The data were cleaned and pre-processed by removing any clear spikes before use as training and testing

data for disaggregation algorithms. Measurement errors from the device were recorded as “999” and therefore could be removed easily. The team acknowledges that these measurement constraints likely hindered the algorithms’ accuracy in disaggregation.

As previously noted, data for test sites were unused for training any algorithms so that it could be exclusively used for measuring accuracy.

Overall CV’s for a few example end uses are presented in Table 1. As described in the Evaluation Metric section, when the NILM method provides a lower CV than logged, it suggests that an evaluator would benefit from using NILM instead of a simple mean when estimating the end use load for a given time period. (CV’s are not presented for the open-source algorithm; this is discussed in the Open-Source Algorithm section below.)

Table 1. Overall CV metric for selected end uses measured for test sites with the end use.

End Use	Logged CV	Proprietary CV	Econometric CV
Central air conditioning	0.68	0.73	0.61
Pool pump	1.37	N/A	3.16
DHW	0.34	0.23	4.26

The overall CV’s provide some general indication of each algorithm’s performance in estimating load shapes as compared to using a simple mean. Results range widely depending on the end use. In the case of central air conditioning, the econometric model offers a slight improvement compared to the logged CV, but the proprietary algorithm does not. The econometric model does not perform well on other end uses such as pool pumps or electric domestic hot water.

The bar plots presented in Figure 3 for central air conditioning loads show both the mean energy use (kWh, first and third row of plot) and the mean CV for different periods. The left column of plots include data from all sites while the right column of plots include data from the test sites only (again, test sites are sites that the algorithm has not previously seen). The team used these bar plots for each end use as its primary diagnostic tool when considering the performance of the algorithms.

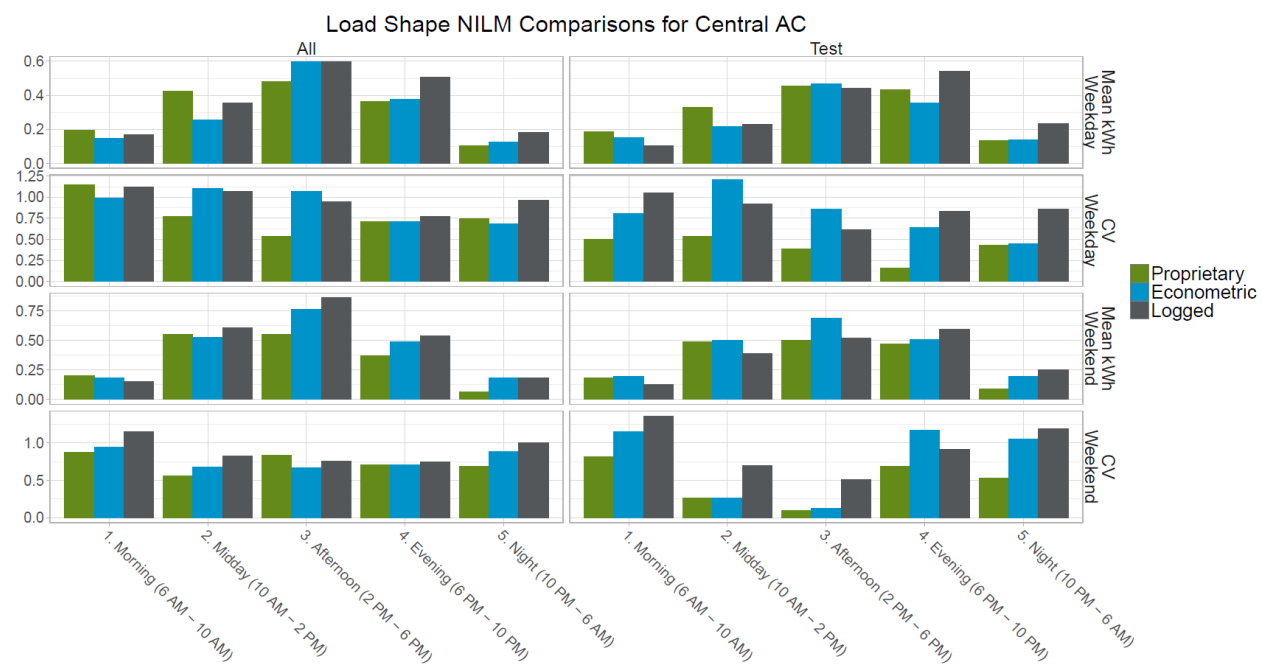


Figure 3. Mean usage (kWh) and mean CV’s for different periods, for all sites in the left column and test sites only in the right column.

Figures 4 and 5 below present the average disaggregated load shapes for all sites with central air conditioning (Figure 4) and pool pumps (Figure 5).

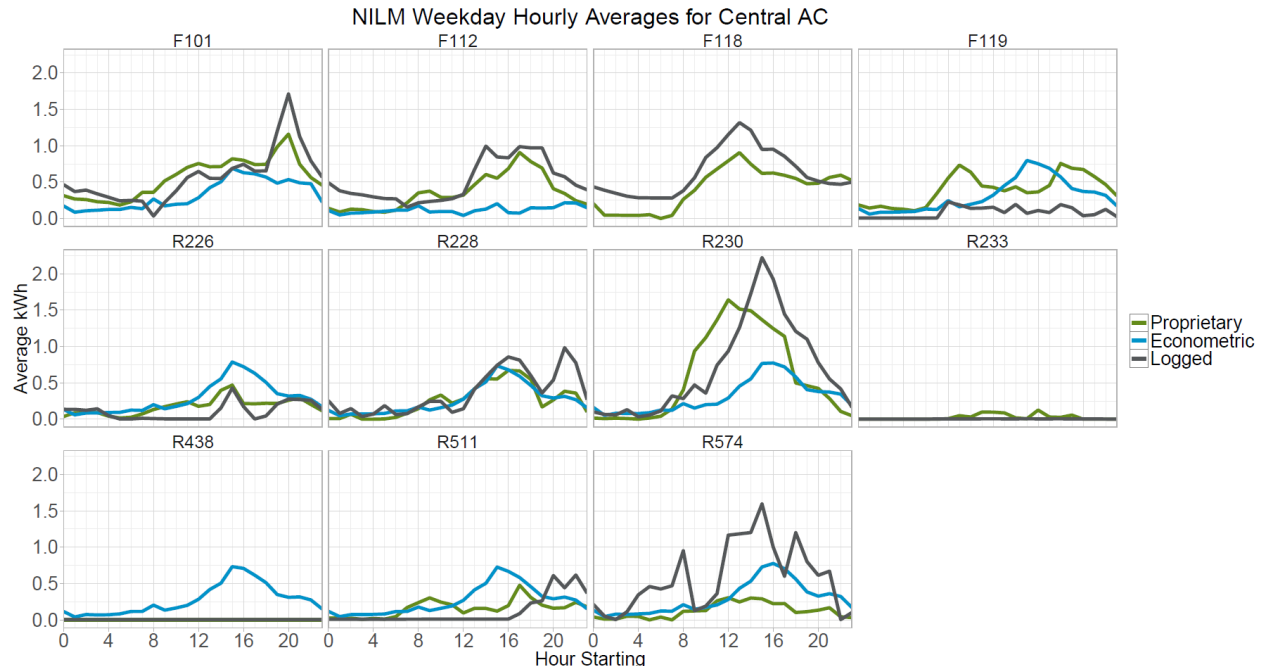


Figure 4. Average hourly time series of air conditioning usage for individual sites.

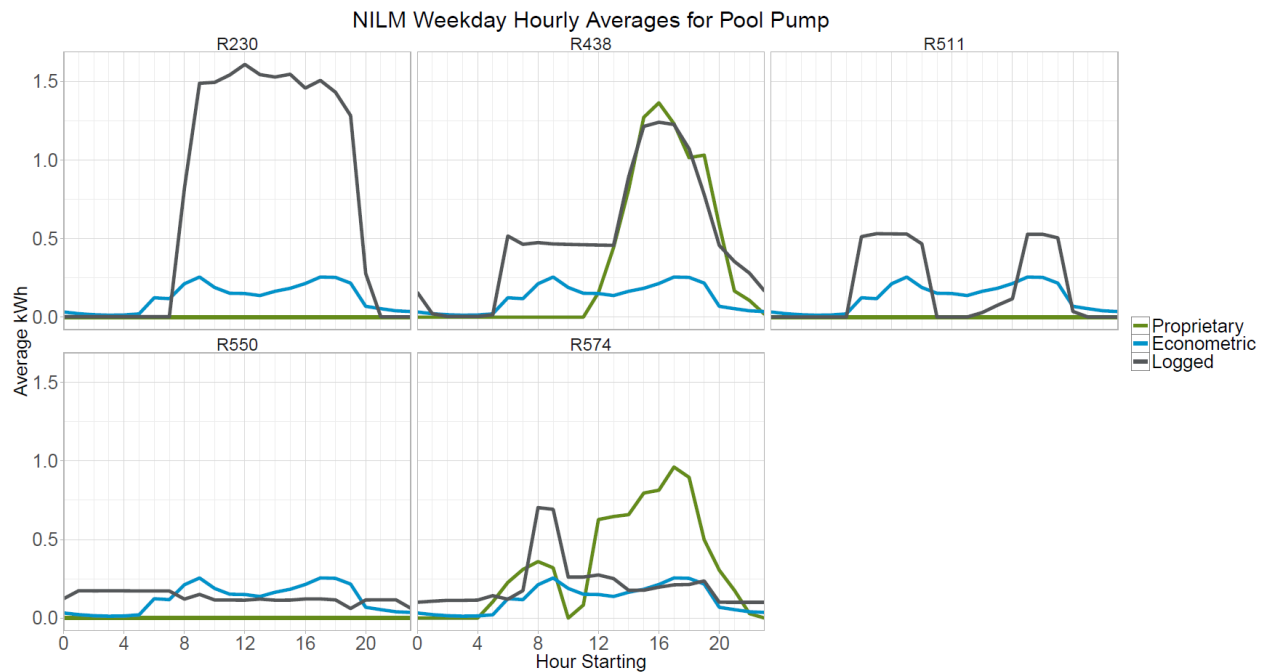


Figure 5. Average hourly time series of pool pump usage for individual sites.

Proprietary Algorithm

When considering the test site plots in Figure 3, the proprietary algorithm appears to both reasonably match the logged mean energy use and reduce the CV, suggesting that central air conditioning could be successfully replaced with NILM as opposed to using an approximation with mean logged values. Central air conditioning is generally the most easily disaggregated end load. Average weekday hourly time series for individual sites are presented in Figure 4.

As can be seen in Figure 5 above, the proprietary algorithm has trouble discerning pool pump load shapes. Pool pumps are of keen interest due to their potentially high usage during the summer period, and therefore performance on this end use is important.

The proprietary algorithm's performance for other end uses was mediocre. Though it was successful for central air conditioning, disaggregating only one end use accurately would not have provided enough cost savings or simplification for a study that needed to develop load shapes for a wide range of end uses.

Open-Source Algorithm

When the open-source algorithm is first trained on the same house it is going to disaggregate (i.e. submetered data for the house is first used to calibrate the model parameters) the team found reasonable performance. However, the evaluation team encountered a well-known problem in the field of NILM for this algorithm, that of lack of generalizability. When the algorithm is tasked with disaggregating a whole-house load for a house that it has not been trained on first, it is unable to recognize the patterns in appliance behavior specific to that household. Figure 6 below illustrates why this is the case: different appliances in the same category can have vastly different power draw profiles.

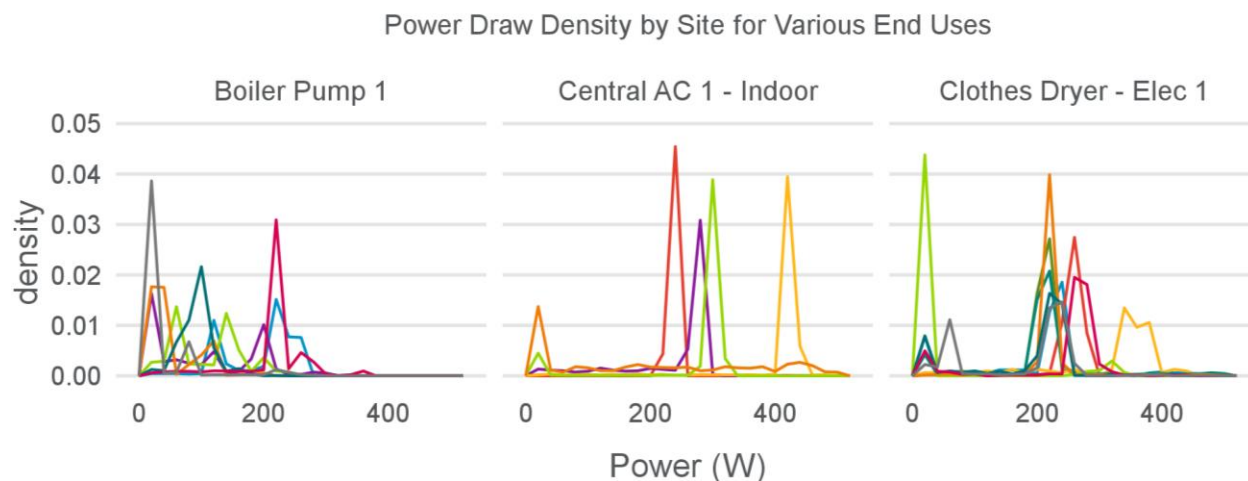


Figure 6. Distribution of power draws (x-axis) for boiler pumps, central AC indoor units, and electric clothes dryers at various houses (colors). The makes and models of appliances vary widely at different homes and accordingly power draws vary, making it difficult to train a NILM algorithm on one set of homes and disaggregate in homes it has not previously seen.

Because of this limitation this algorithm the evaluation team did not consider the open-source algorithm a viable candidate for use in the baseline study, where the algorithm would be expected to disaggregate households it had not been trained on, and therefore the team did not pursue complete evaluation of its accuracy.

Econometric Model

In contrast to the open-source algorithm, the econometric model performed moderately well when looking at the average load shapes across all households in the study. However, the accuracy of the econometric model was insufficient when considering specific household results. The econometric model cannot predict individual loadshapes for specific households; the model derives an hourly load shape for a group of households from the variation in appliance saturation across those households and thus results only in an average load shape for the group of households.

Figure 7 shows the predicted average appliance load shapes compared with the average hourly load across households. The colored bands are the predicted load shapes from the econometric model for each end use. The black line is the average metered hourly whole house load. If the model is accurately predicting end-use loads, the gap between the aggregated predicted load shapes and the whole house load should only represent the un-modeled end-uses (lighting, consumer electronics, or other appliances without variation in ownership). While the aggregated appliance loads follow the general household load shape, the peaks do not line up and there is a significant portion of the load that remains unaccounted for. Although not all end-uses were modeled, the gap is larger than would be expected to result from these end uses and is therefore unexplained.

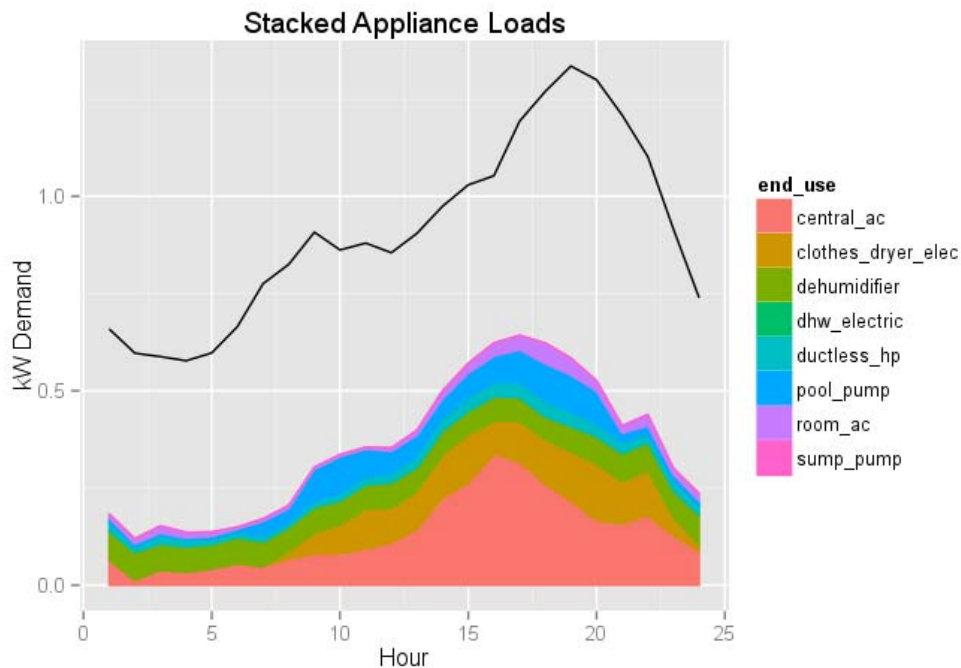


Figure 7. Average hourly disaggregated loads in econometric model. There is a larger-than-anticipated gap between the sum of the disaggregated loads and the whole-house loads that cannot be fully explained by lighting, consumer electronics, and various other end-uses for which there was insufficient saturation for modeling purposes.

When looking at the disaggregated loads at the household level, the econometric model did not perform well. Figures 4 and 5 shows the predicted central air conditioning and pool pump loads by household. While there are a couple sites where the end use loads shapes appear reasonable, most do not line up with the household load profiles.

Mean energy use and CV for central air conditioning loads by period-of-day and weekday/weekend bins for the econometric model are also presented in Figure 3. During periods of increased AC use (afternoon and evening), both the proprietary algorithm and the econometric model provide reasonable approximations of total energy use and reduced CV for the proprietary algorithm

compared to the logged CV suggest that its disaggregation estimate may be useful. However, the econometric model CV does not offer accuracy improvements during weekday afternoons, a critical period for utilities.

Conclusions

While software may be able to disaggregate load shapes for some end-uses such as cooling or pool pumps, algorithms are not yet able to accurately discern loadshapes for end-uses that are not weather-correlated or are entirely driven by stochastic human behavior. These problems are overcome with training periods; some consumer-facing algorithms and products that display home energy use ask homeowners first to actively label end uses as they occur by switching appliances on and off one by one so that the algorithm can “learn” what appliance behavior looks like specifically in that house, enabling more reliable estimates of end-use patterns for homeowners. However, for utilities or program evaluators wishing to understand load shapes in a large number of homes at low-cost and low-invasiveness to the household, this need for individual training is not desirable. Recent developments in deep learning, advanced machine learning algorithms that require a lot of data, may be a path forward. Indeed, Jack Kelly and William Knottenbelt suggest that deep neural networks for energy disaggregation perform better than previous algorithms on houses that have not been previously seen (Kelly 2015). In Decker 2017, the evaluation team considers the cost-benefit trade-offs of using NILM technologies further.

References

- Blaney, John, Mark Inglis, and Asa Janney 1994. “Hourly Conditional Demand Analysis of Residential Electricity Use.” *ACEEE Proceedings, 1994 Summer Study on Energy Efficiency in Buildings*, Vol. 7. <http://aceee.org/proceedings>
- Caves, Douglas, Joseph Herriges, Kenneth Train, and Robert Windle 1987. “A Bayesian Approach to Combining Conditional Demand and Engineering Models of Electricity Usage.” *The Review of Economics and Statistics*, Vol. 69 No. 3: 438-448.
- Kelly, Jack and William Knottenbelt 2015. “Neural NILM: Deep Neural Networks Applied to Energy Disaggregation.” September 2015. <https://arxiv.org/pdf/1507.06594.pdf>
- Makonin, Stephen and Bob Gill 2015. “Exploiting HMM Sparsity to Perform Online Real-Time Nonintrusive Load Monitoring (NILM).” *IEEE Transactions on Smart Grid*.
- Parti, Cynthia and Michael Parti 1980. “The Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector.” *The Bell Journal of Economics*, Vol. 11 No. 1: 309-321.