Load Disaggregation Technologies: Real World and Laboratory Performance

Ebony T. Mayhorn, Pacific Northwest National Laboratory
Gregory P. Sullivan, Efficiency Solutions
Joseph Petersen, Ryan S. Butner and Erica M. Johnson, Pacific Northwest National Laboratory

ABSTRACT

Over the past ten years, improvements in low-cost interval metering and communication technology have enabled load disaggregation through non-intrusive load monitoring (NILM) technologies, which estimate and report energy consumption of individual end-use loads. Given the appropriate performance characteristics, these technologies have the potential to enable many utility and customer facing applications. However, there has been skepticism concerning the ability of load disaggregation products to accurately identify and estimate energy consumption of end-uses; which has hindered wide-spread market adoption. A contributing factor is that common test methods and metrics are unavailable to evaluate performance without conducting large-scale field demonstrations and pilots, which can be costly. Without common and cost-effective methods of evaluation, advanced NILM technologies will continue to be slow to market and potential users will remain uncertain about their capabilities.

This paper reviews recent field studies and laboratory tests of NILM technologies. Several important factors are identified for consideration in test protocols so their results reflect real world performance. Potential metrics are examined to highlight their effectiveness in quantifying disaggregation performance. This analysis is then used to suggest performance metrics that are meaningful and of value to potential users and that will enable researchers/developers to identify beneficial ways to improve their technologies.

Introduction

Access to granular energy consumption information of individual end-use appliances and equipment can improve the energy efficiency of buildings and resiliency of the electric power grid. This information has the potential to enable many use cases (i.e. ways a user may apply the information collected), such as follows:

- Inform residential customers, utilities, and transmission /distribution resource planners about energy use and load shapes of individual electrical appliances.
- Enable M&V for utility demand response (verify individual electrical appliances have modified consumption accordingly in response to demand response command) and efficiency programs (verify savings from efficiency measures via more efficient end-use technologies).
- Enable diagnostics and preventative maintenance of electrical appliances/systems for improved building operations.

However, current methods to monitor individual end uses are either very expensive and intrusive to the building (e.g. directly sub-metering end uses) or indirect and non-intrusive, which are unproven thus far to perform at the levels desired for relevant use cases (Mayhorn et al. 2015, Armel et al, 2012, Zoha et al. 2012, Zeifman et al., 2011a, Berges et al., 2010).

There are several types of indirect monitoring methods to acquire end-use energy data. Non-intrusive load monitoring (NILM) technologies have the potential to offer a low-cost, scalable method for acquiring end-use data. NILM is a data analytics approach that uses limited measurements taken at the whole building level (e.g. utility meter or current transformers installed at the building breaker panel) coupled with assumptions about behavioral habits (i.e. average load shapes) to disaggregate end-use energy consumption. Since buildings have a wide spectrum of end-use types, each with an array of models and unique behavioral patterns, NILM vendors face significant challenges with identifying and inferring energy consumption of end uses.

Although NILM has the potential to enable many use cases that support energy efficiency, there are market barriers preventing widespread adoption of this technology. First, expensive large-scale field studies, which can take more than a year to complete (Mayhorn et al. 2015, Pecan Street 2015, White 2014) due to the need to capture load diversity, seasonal loads, and complete the analysis, are required to demonstrate and understand the performance of NILM products in the field. For NILM technologies to be successful, their target customers need a valid demonstration of performance to justify investment in the products.

Second, the NILM industry does not have a clear understanding from potential users which capabilities and performance characteristics are important to consider for the range of use cases these technologies could potentially support. For example, use cases may involve a different set of appliances than those desired by potential users. Without clear understanding of the range of use cases and expectations for those use cases, it is difficult for the NILM industry and researchers to prioritize and focus R&D efforts to develop the technology to have the capabilities and performance desired.

Third, previous and/or current evaluations of NILM have led to uncertainty in the market and raised skepticism about its performance capabilities. Several NILM lab test and field study performance evaluations have been performed and published recently (Butner et al. Dec 2013, EPRI 2013, Mayhorn et al. 2015, Pecan Street 2015, White 2014), as summarized in Table 1. Each study considers unique appliance types, evaluation periods, metrics, sample datasets or tests, time scales or intervals for applying metrics, and evaluation methods.

Increasing confidence in NILM disaggregation capabilities is needed to accelerate market adoption. This calls for the NILM industry to come to an agreement on the appropriate test protocols. For this reason, PNNL has initiated an effort to develop test protocols to evaluate NILM performance in collaboration with the NILM Protocol Development Advisory Group formed. The advisory group consists of 17 members, representing key stakeholders, including NILM vendors, researchers, field evaluators as well as utility/energy efficiency organizations and alliances. The goal of the test protocols is to mitigate the need for repeated large-scale field studies and eliminate the confusion as to which metric(s) are necessary to objectively quantify, compare and communicate NILM performance. The test protocols should have the appropriate elements necessary to reflect real-world performance. In addition, metrics should be selected and applied carefully so results are not misinterpreted. This paper discusses the critical factors to

consider when developing a test protocol that captures real-world performance. It also examines a set of candidate NILM performance metrics by applying the metrics to several scenarios representing different types of end-use behaviors and possible types of error in NILM energy estimates. Recommendations are given for suitable metrics, which have also been agreed on by the NILM Protocol Development Advisory Group.

Table 1. Overview of NILM performance evaluations

	End use types	Evaluation Period	Metrics	# of tests / samples	Time scales	Evaluation Method
EPRI Round 1 Lab Test	HVAC; Water heater; Clothes dryer; Range; Refrigerator; Freezer; Pool pump; Fan; Lighting; Microwave;	1 week training period; 2 week evaluation period	RMSD; μ, σ; average % estimated premise energy use to total; # of isolated loads	1 test	1 min 5 min 15 min 1 hr Daily Weekly	Laboratory test; emulated typical use of average single-family home; appliance schedules based on DOE Building America benchmark
PNNL 2013 PNNL Lab Homes Test	25 W table lamps; 240 W hardwired light fixtures; 2 kW Electric resistance water heater (ERWH)	1 week	# of correctly identified loads; event detection accuracy; energy estimation accuracy; repeatability	1 test	Weekly	PNNL Lab Homes; simultaneous / sequential loads scheduled at varying periods (1 min, 10 min, 1 hour)
Pecan Street/ EEme Assessment 2015	HVAC; Refrigerator; Clothes dryer; Dishwasher	1 year	Absolute error; relative error	264-homes	15 min	Data-driven evaluation; metered energy data of actual homes
NEEA NILM Field Study	Clothes dryer; Water heater; Refrigerator; Furnace; ER Heat; Freezer; Dishwasher; Oven/Range; Clothes washer	24 weeks filtered data	% standard deviation explained; r-squared	4 homes	5 min or Daily	NILMs installed in different PNW region homes; each home contained different set of appliances types/models
SDG&E NILM Case study	Electric Vehicle; HVAC; Pool pump; Oven; Refrigeration/ Freezer; ERWH; Dryer	1 year filtered data	RMSD; μ, σ; f-score; error in assigned power; r-squared; error in total assigned energy	11 homes	Daily or monthly	10 sec energy use data collected using Rainforest Automation Eagle gateways or 1 hr SDG&E Green Button Connect data

Important Elements of Test Protocol to Assess NILM Performance

The objective is for this protocol is to overcome some of the barriers with developing the technology to the desired levels for priority use cases. Specifically, the goals of the test protocol are to: (1) minimize the need for large-scale field studies to demonstrate and understand performance of NILM, and (2) improve confidence in NILM performance capabilities through a common evaluation method. Several lessons learned have been identified and itemized based on examinations of NILM performance and investigations into current NILM product specifications. Each of these lessons should be considered when making decisions regarding particular elements to include in a test protocol that is able to capture real-world performance.

Lessons

Lesson 1: NILM industry was unclear about desired performance requirements for a wide range of potential use cases. Performance requirements haven't been clearly defined for the wide range of use cases that NILM has the potential to support or enable. Therefore, a common set of important performance characteristics (i.e. NILM can identify end uses properly, detect events, estimate energy use, etc.) should be communicated to the NILM market in a uniform way, similar to how a set of nutrition facts are presented on food labels for a wide range of consumer types to make food choices. Communicating the information in a clear, common way would reduce uncertainty in the market about NILM performance as well as enable potential NILM users to make decisions about whether NILM is suitable for their use cases. When developing test protocols, it is imperative to understand and define a general list of performance characteristics that are important to measure and communicate to the NILM market. This list should be considered when determining the appropriate set of protocols and metrics to evaluate performance characteristics. In addition, there is a large and emerging spectrum of end-use types available in buildings. Every end use cannot be considered in assessments of NILM performance in the near term. Therefore, an understanding of the end-use types that are of interest for high priority use cases should be used to determine the most suitable set of end uses to include in the evaluation protocols.

Lesson 2: May take measurement inputs at the different sampling rates. NILM products under development and/or commercially available require building level measurement inputs to be taken at sampling rates that range from megahertz (MHz) to hourly. For approaches requiring building measurement inputs at one-minute time intervals or greater, a data-driven test method could be chosen using the metered data collected from field studies. A data-driven methodology would use whole building and sub-metered building energy data that have been previously collected (for example the ~1,200 occupied residential and commercial buildings data set collected by Pecan Street at as low as one-minute intervals over a year). The whole-building energy data could be used as inputs to NILM products under evaluation. Then, the energy estimates reported for each end use to be compared to the corresponding sub-metered end-use data. However, any NILM product requiring inputs at less than one-minute intervals would be ineligible for testing using a data-driven test method. Therefore, a laboratory test method is another option to ensure a fair comparison across all NILM products regardless of input sampling rate. A laboratory test would be more expensive than a data-driven protocol due to the

costs of purchasing a representative set of specific end-uses. There is a wide spectrum of end-use types and models available that have different programmed modes for use, so selecting a proper set of end uses to include in the test and base the assessment on would be a critical decision to ensure the test is able to characterize real world performance.

Lesson 3: May rely on load pattern library and/or other behavioral cues. NILM algorithms are usually proprietary, but many rely on assumptions about end-use behavior, such as load shape patterns, to inform energy use estimates. Therefore, NILM evaluations should incorporate realistic or typical use behavior of end uses. In particular, for each end-use type selected for evaluation of NILM, a statistically large number of tests or samples data sets should be selected to capture real-world performance.

Lesson 4: May require a training/learning period. The length of experiments to evaluate NILM performance should be selected carefully. NILM technologies may need training to learn end use signatures, so NILM performance is not unfairly penalized during those periods. Test protocols should include a separate training period for seasonal end uses (i.e. space cooling and heating equipment).

Lesson 5: May have limitations on time intervals for reporting energy use. NILM may not be able to evaluate all desired performance characteristics identified because of limitations on the time intervals for reporting energy use (i.e. 1 hr or 24 hr intervals). For example, users may want to understand performance at short to long time scales or understand performance in detecting ON/OFF events of end uses. Therefore, for all performance characteristics to be quantified, it is important to indicate time intervals in the reported results when applying the metrics to evaluate the relevant performance characteristics.

Lesson 6: NILM products have labeling inconsistencies. While some efforts have been initiated (Kelly et al. 2014), a naming convention for labeling end use types has not been fully adopted by the NILM industry. Without consistent naming conventions, it becomes challenging to evaluate NILM performance by end use. For example, assume the compressor and defrost elements of the same refrigerator were tracked by a NILM, but are labeled as two separate loads (i.e. Load 1 and Load 2). If the Load 1 NILM data is selected, instead of both Load 1 and 2 data, the NILM performance results will be very different. Also, NILM products may track some end uses well but label the end uses incorrectly. Test protocols should be defined to consider labeling or identification inconsistencies so that the performance can be uniformly assessed.

Lesson 7: Performance can be misunderstood if the appropriate metrics are not selected and applied carefully. Based on data collected from the Northwest Energy Efficiency Alliance (NEEA) NILM Field Study (Mayhorn et al. 2015), the "accuracy" in a NILM energy use estimates was computed to be 72.4% for a refrigerator, based on a 24-week study period and the relative error metric. Figure 1 shows the NILM and baseline energy-use profiles for the refrigerator in the home over a 24-hour period. From this figure, it is clear that the NILM product is not tracking the energy use of a refrigerator even though the energy accuracy was determined to be very high. In this case, the NILM device was actually reasonably tracking the energy-use profile of a water heater. Because the water heater has several large energy draws over short

durations that are comparable to the total energy consumption of a refrigerator over a longer duration, the results based on this metric and how it was applied could be misinterpreted if it were used to convey the NILM device performance.
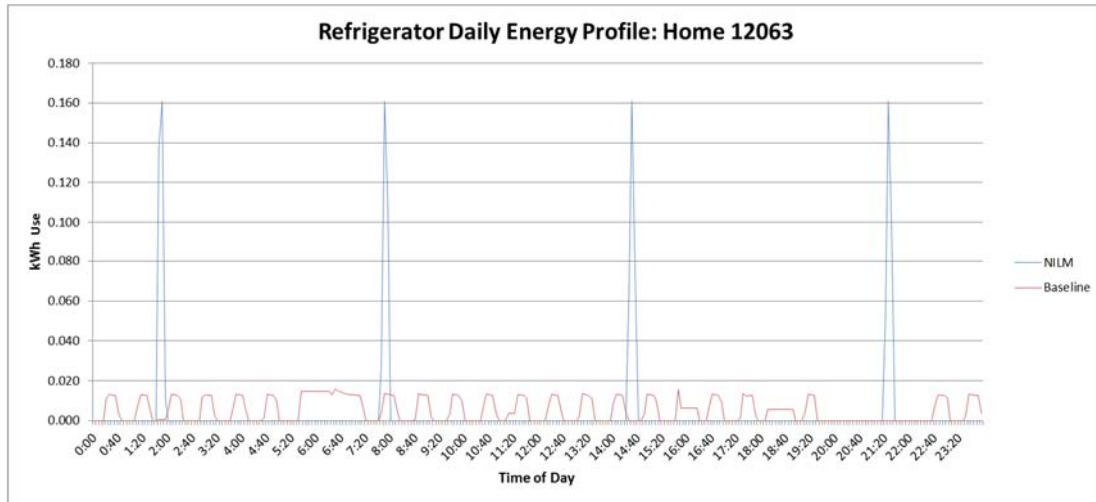

Figure 1. Refrigerator Energy Profile (over 1 day) from a home in of NEEA NILM Field Study

Candidate NILM Performance Metrics

Based on current research (Butner et al. 2013a, Mayhorn et al. 2015, Zeifman et al., 2011b, Liang et al., 2010), the NILM industry has yet to select and adopt a common set of metrics for performance verification of NILM devices. The conclusion drawn from the initial NILM research was that there are many aspects that should be considered when defining the "accuracy" or performance of a NILM, and multiple metrics may be required to evaluate performance. A list of candidate metrics were assembled from previous NILM research (Batra et al. 2014, Holmes 2014, Mayhorn et al. 2015, Pecan Street 2015, Timmermans and Sachs 2015), as well as from metrics proposed by the Advisory Group formed. The metrics fall into the following two main categories that may be of interest to potential users of NILM:

- Event Detection (ED) Performance – metrics designed to evaluate a NILM's ability to track the energy use patterns of end-uses over time. Using these metrics, the performance is quantified with consideration for the number of true positive, false positive, true negative and/or false negative events computed using disaggregated energy use data.
- Energy Estimation (EE) Performance – metrics designed to characterize and evaluate NILM disaggregated energy use versus known "ground truth" end-use data. This set of metrics consist of basic statistics (e.g. mean and standard deviation of error), advanced statistics indicating goodness of fit of NILM and ground truth data (e.g. R-squared and percent standard deviation explained) and other error statistics evaluating relative, average, and per-event/time step error.

Even though ED metrics indicate NILM's ability to codify individual events and track energy use patterns overtime, the NILM Protocol Development Advisory Group came to an agreement that ED performance was implicit in the EE metrics considered, so only the EE performance metric category is needed. Table 2 presents the full list of EE performance metrics evaluated.

Table 2. Energy estimation performance metrics evaluated

Metric	Metric Equation
EE 1: Relative Error	$$\frac{\sum_{i=1}^{N} E_i - \sum_{i=1}^{N} \hat{E}_i}{\sum_{i=1}^{N} E_i}$$
EE 2: Root Mean Square Deviation	$$1 - \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(E_i - \hat{E}_i)^2}}{\bar{E}}$$
EE 3.1: Average Error	$$\frac{1}{N}\sum_{i=1}^{N} \Delta E_i$$
EE 3.2: Standard Deviation of Error	$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta E_i - \overline{\Delta E})^2}$$
EE 4: Percent Standard Deviation Explained	$$1 - \sqrt{1 - r^2}$$ $$r^2 = 1 - \sum_{i=1}^{N}\frac{(E_i - \hat{E}_i)^2}{(E_i - \bar{E})^2}$$
EE 5: R Squared	$$r^2 = 1 - \sum_{i=1}^{N}\frac{(E_i - \hat{E}_i)^2}{(E_i - \bar{E})^2}$$
EE 6: Energy Error	$$\frac{\sum_{i=1}^{N}\left|\hat{E}_i - E_i\right|}{\sum_{i=1}^{N} E_i}$$
EE 7: Energy Accuracy	$$e^{-\alpha(Energy\ Error)}$$
EE 8: Match Rate	$$\frac{\sum_{i=1}^{N} \min\{E_i, \hat{E}_i\}}{\sum_{i=1}^{N} \max\{E_i, \hat{E}_i\}}$$
EE 9: Mean Standard Error	$$\frac{\sigma}{\sqrt{N}}$$

$\hat{E}_i$= NILM energy data at each time interval i σ = standard deviation of the error over the data set
E_i= Metered data ate each time interval i α = mapping factor defined to be 1.4
$\bar{E}$ = Average metered energy over the data set. .   i = indexing variable for each time step
ΔE = error between the NILM and metered data
N = number of observation or events based on each time step

Metrics Evaluation and Recommendations

NILM energy use estimates can embody multiple types of error. For example, NILM data may contain a combination of under-estimated energy use, time offsets in estimates as well as

missed and falsely detected events. Therefore, it is challenging to determine whether the values obtained from applying metrics accurately represent NILM performance. To recommend intuitive and meaningful metrics that are suitable for NILM performance verification, the response of all 10 candidate metrics in Table 2 were examined under a variety of hypothetical and error scenarios, each with different error types and magnitudes in the data, as well as different types of appliances. Specifically, the metrics were applied to a set of real-world appliance energy use data that were collected as part of the NEEA NILM field study (Mayhorn et al., 2015). These real-world data, 5-minute interval data by end-use were duplicated and altered to systematically to introduce known errors in NILM energy use estimates. The discrepancies introduced between the NILM estimates and real world data were intended to model typical errors encountered with NILMs and are based on findings of the NEEA RBSA field study (Mayhorn et al. 2015). The types of errors considered are:

- Consistently over-/under- estimating energy by a nominal percentage (e.g., 5%, 25%, 50%) of actual appliance energy use.
- Consistently over-/under- estimating energy by a constant value relative to actual appliance energy use.
- False positives/negatives – estimates include percentage of false positive/negative events relative to total number of actual ON events.
- Constant time offsets in runtimes – energy estimates are delayed by one or more time intervals.

To better illustrate some of the real-world data errors identified, Figure 2 below presents a single day of actual NILM and metered refrigerator energy use data from the previous research. Of interest is the consistently lower energy reading detected by the NILM and the missed events related to the refrigerator defrost cycle occurring at about 6:00 AM. Data such as these, with the discrepancies, were useful in identifying and quantifying data errors and became the basis for the error scenarios considered.
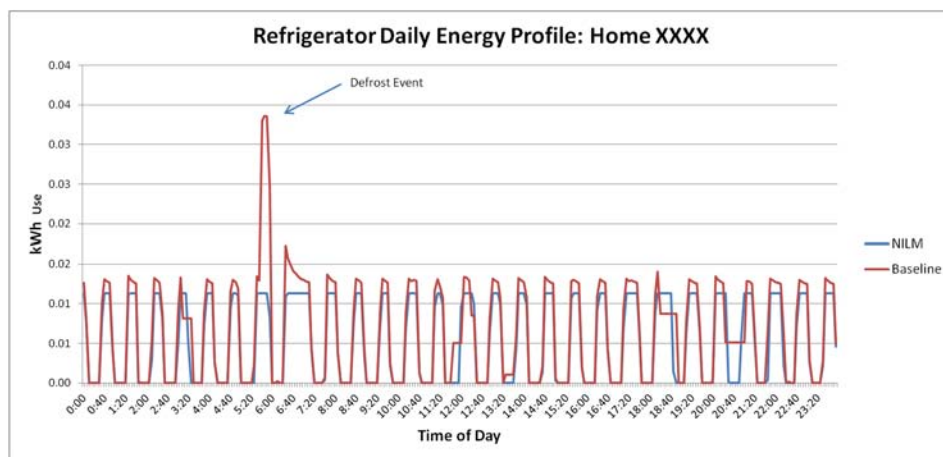

Figure 2. Refrigerator Daily Energy Profile Reporting Baseline (Metered) and NILM Energy Use

Table 3 presents the list of NILM error scenarios generated from real world energy use data for three different appliance types: refrigerators, clothes dryers, and water heaters (Mayhorn et al. 2015). Some of these errors were exaggerated to test the limits of the metrics and identify non-linear responses among other metric issues. The data sets for each simulated error scenario are one week in duration and constructed with five-minute intervals.

Table 3. Simulated Error Scenarios to Evaluate NILM Metrics

NILM Error	Error Scenario	Data Modifications
NE 1	One interval offset of actual appliance energy data	NILM data shifted forward by 1 metering interval
NE 2	Five interval shift of data	NILM data shifted forward by 5 metering intervals
NE 3	NILM under-estimates actual energy by 5%	All NILM data values reduced by 5%
NE 4	NILM under-estimates actual energy by 25%	All NILM data values reduced by 25%
NE 5	NILM under-estimates actual energy by 50%	All NILM data values reduced by 50%
NE 6	NILM over-estimates actual energy by 25%	All NILM data values increased by 25%
NE 7	NILM estimates energy perfectly but include 5% missed events	5% of events removed from NILM data set
NE 8	NILM estimates energy perfectly but include 25% missed events	25% of events removed from NILM data set
NE 9	NILM estimates energy perfectly but include 50% missed events	50% of events removed from NILM data set
NE 10	NILM estimates energy perfectly but include 25% false events	NILM data set has 25% additional false events introduced
NE 11	NILM over-estimates energy by a low constant value	All NILM data values increased by 0.001 kWh
NE 12	NILM over-estimates energy by a high by a constant value	All NILM data values increased by 0.004 kWh

Metric Evaluation Approach

The metric evaluation process was designed to apply the proposed metrics in a consistent manner across all simulated error scenarios and then evaluate the response to each scenario. For each scenario, each metric was applied in three different ways to understand sensitivity to the number of true negative events encountered as well as to the time period considered. In the first, all data points over the full one week data sets were used in the evaluation. Second, all records of true negative events were omitted. The third way is similar to the first, but only takes a half week of data into account. The analysis was automated using MATLAB resulting in compiled output tables suitable for results comparison.

To evaluate the metrics, the resulting values from all scenarios were reviewed against the expected performance, and the metrics were assigned a rating for each scenario on a scale of 0 - 3, where: "0" indicates that the metric value is not apparent or explainable, "1" indicates an

apparent trend in metric value but not a value near the expected, "2" indicates the resulting metric value approximates the value expected, and "3" indicates the resulting metric value is precise based on value expected. To decide the rating, several factors were kept in mind: (1) precision of fit to known error, (2) occurrence of values outside of 0 to 1 range, (3) occurrence of non-linearity of values with the progression of error across scenarios (e.g. 5%, 25% and 50% error), and (4) intuitive and explainable results. Once these ratings were determined based on each error scenario, a score for each metric was assigned by summing the ratings to compare metrics. Comparing all metrics in this fashion allowed the evaluation team to identify those metrics which are most robust to a wide range of error types and most suitable for assessing NILM performance.

Metric Evaluation Results

When evaluating the energy estimation metrics, all scenarios listed in Table 3 were considered. Therefore, the highest possible score for this set of metrics is 36. After reviewing the values for the metrics evaluated against the expected outcomes, it became clear that a number of the metrics operated well with simple errors (e.g., a 5% error in NILM data, NE 3 in Table 3), but exhibited out-of-range (negative values or values greater than 1) or did not scale as expected as the errors were systematically increased (e.g., the progression of a 5% error to 25% and 50%). The tallied scores for each metric evaluated are given in Table 4. EE 6, EE 7 and EE 8 were the three metrics that stood out from the rest. For all error scenarios, the values obtained were close to the performance values expected. However, in some scenarios with relatively large error, EE 6 would result in values greater than 1, making the metric values less intuitive and explainable. The EE 7 metric was proposed with the intent to ensure that the quantified performance values are between 0 and 1. However, this metric would require tuning of the α parameter to ensure the specific values accurately represent the performance. EE 8 does not require tuning of parameters to ensure values are not misinterpreted, and the resulting values based on each scenario appear to be an accurate representation of NILM performance.

Table 4. NILM metric scores for candidate EE performance metrics

Metric	Metric Description	Metric Score
EE 1	Relative Error	30
EE 2	Root Mean Square Deviation	16
EE 3.1	Average Error	12
EE 3.2	Standard Deviation of Error	12
EE 4	Percent Deviation Explained	16
EE 5	R Squared	12
EE 6	Energy Error	32
EE 7	Energy Accuracy	30
EE 8	Match Rate	35
EE 9	Mean Standard Error	12

Metric Recommendations and Application

Based on the metrics analysis completed, it is recommended that EE 8 be used to assess energy estimation performance of NILM. The analysis and results presented in this paper were discussed with the NILM Protocol Development Advisory Group, and a consensus was reached that the EE 8 should be used for NILM performance evaluations. However, it is important for metrics to be applied carefully such that performance values have the same meaning across different end uses of various sizes (high/low energy use loads) and patterns of use (regular/intermittent load cycles). For example, the errors computed relative to low energy use may be less significant when compared to a high energy use load. In addition, some end uses operate less frequently than others, so metric results may not be representative of performance if a statistically significant number of run times are not captured. Given these issues, it is recommended that any metric evaluation takes into account some leveling of energy use or other basis for fair performance comparisons across different end uses. This basis may be in the form of a fixed energy use for each NILM evaluation per end use (e.g., a 10 kWh actual energy for refrigerator and a dryer) for EE metrics, a fixed number of actual events per end use for ED metrics, or a fixed duration that is able to capture a range of conditions (e.g., 1 week of a typical operation patterns).

Conclusions

To overcome market barriers for NILM technologies, it is critical that a common set of test protocols and metrics be developed to convey performance in a consistent and meaningful way to potential users as well as to reduce the need for large scale field studies to evaluate performance. Several important factors were offered for consideration when developing test protocols to reflect real world performance. Thus far, as part of the NILM protocol development effort at PNNL, 10 candidate metrics have been identified and examined from the literature and advisory group for assessing energy estimation performance. The metrics were applied to a variety of error scenarios with known performance expectations, each representing a different type/magnitude of error that can be encountered by NILM. These scenarios were created to determine whether the responses of the metrics are intuitive. Metric EE 8 was demonstrated as the most robust in representing performance across all error scenarios considered and is recommended for evaluating NILM performance. The value of this metric is bounded between 0 and 1, therefore could be used to easily convey performance. In addition, all values reported were very close to the performance expectations in all error scenarios considered. The next steps in the PNNL protocol development effort are to weigh the advantages and disadvantages of developing a data driven versus laboratory test method for evaluation, and then iterate with the Protocol Development Advisory Group to develop the appropriate protocols for agreement by important NILM stakeholders.

References

Armel, C., Gupta, A., Shrimali, G., and Albert, A. "Is Disaggregation The Holy Grail of Energy Efficiency? The Case of Electricity". Precourt Energy Efficiency Center, Stanford University, Palo Alto CA.

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

Berges ME and E Goldman. 2010. "Enhancing Electricity Audits in Residential Buildings with Nonintrusive Load Monitoring." *Journal of Industrial Ecology, Environmental Applications of Information & Communication Technology* 14(5): 844-858. Carnegie Mellon University, Pittsburgh, Pennsylvania.

Butner, RS. 2013. *Non-Intrusive Load Monitoring Assessment: Literature Review and Laboratory Protocol.* Accessed October 2013. Richland, WA: Pacific Northwest National Laboratory. http://www.pnnl.gov/main/publications/external/technical_reports/PNNL-22635.pdf

Electric Power Research Institute. 2013. *Non-Intrusive Load Monitoring (NILM) Technologies for End-Use Load Disaggregation.* Palo Alto, CA: 2013. 3002001526.

Holmes, C. 2014. *Non-Intrusive Load Monitoring (NILMS) Research Activity: End-User Energy Efficiency & Demand Response.* Presentation at the 2014 AEIC Load Research Analytics Workshop. http://publications.aeic.org/lrc/2014_Workshop_NonInstrusiveApplianceLoadMonitoring.pdf

Kelly, J., and Knottenbelt, W. (2014). *Metadata for Energy Disaggregation*. The 2nd IEEE International Workshop on Consumer Devices and Systems (CDS) in Västerås, Sweden. arXiv:1403.5946

Liang J., Ng S.K.K., Kendall G., Cheng J.W.M. 2010. "Load signature study Part I: Basic concept, structure, and methodology." IEEE Trans. Power Del. 2010; 25:551–560.

Mayhorn, ET. 2015. *Characteristics and Performance of Existing Load Disaggregation Technologies*. Accessed May 2015. Richland, WA: Pacific Northwest National Laboratory. http://www.pnnl.gov/main/publications/external/technical_reports/PNNL-24230.pdf

Pecan Street Smart Grid Demonstration Project. 2015. *Setting the Benchmark for Non-Intrusive Load Monitoring: A Comprehensive Assessment of AMI-based Load Disaggregation.*

Pecan Street Smart Grid Demonstration Project. 2014. *About Pecan Street.* http://www.pecanstreet.org/about/

Timmermans, C. and von Sachs, R., 2015. "A novel semi-distance for measuring dissimilarities of curves with sharp local patterns. " Journal of Statistical Planning and Inference, 160, pp.35-50.

White, B., M. Esser. *Residential Disaggregation – Final Report*. NegaWatt Consulting, Inc. 2014. Accessed February 2016. http://www.etcc-ca.com/sites/default/files/reports/et13sdg1031-residential_disaggregation.pdf

Zeifman M, C Akers, and K Roth 2011a. *Nonintrusive appliance load monitoring (NIALM) for energy control in residential buildings*. Fraunhofer Center for Sustainable Energy Systems, Cambridge, Massachusetts. Accessed May 2016. http://cdn2.hubspot.net/hub/55819/file-2309099257-pdf/docs/eedal-060_zeifmanetal2011.pdf?t=1461937648169

Zeifman M., Roth K. 2011b. "Nonintrusive Appliance Load Monitoring: Review and Outlook." IEEE Trans. Consum. Electron. 2011;57:76–84.

Zoha A, A Gluhak, MA Imran, and S Rajasegarar. 2012. "Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey." *Sensors*. 2012 Vol: 12(12) 16838:16866. Basel, Switzerland. Accessed July 2013. www.mdpi.com/1424-8220/12/12/16838.pdf