Measuring Demand Savings with Smart Thermostat Data

Ethan Goldman and Abiodun Iwayemi, Vermont Energy Investment Corporation, Burlington, VT
Jennifer Robinson, Ram Narayanamurthy, and Ben Clarin, EPRI, Palo Alto, CA
Robert Ruskamp and Marc Shkolnick, Lincoln Electric System, Lincoln, NE

ABSTRACT

There is no question that detailed data streams from connected thermostats offer a comprehensive view of smart thermostat operations and of building operations where the thermostats are installed. But detailed data streams from hundreds or thousands of smart thermostats contain raw, messy data. What does it take to convert these messy data into verifiable measurements of energy use and savings? This paper shows how the Vermont Energy Investment Corporation (VEIC) accurately calculated both demand response and energy efficiency savings with smart thermostat data—and validated the method. We also present unexpected insights from this approach, and lessons learned about working with thermostat data.

This paper addresses the challenges and benefits of working with device-level interval data, and proposes some tactics for successfully extracting value from high-volume data. The analysis revealed data quality issues that needed attention, and exposed problems with certain applications of demand response HVAC control strategies. Spotlighting these issues led to fixes in the ways the vendor called the events, and to more nuanced evaluation of the impacts. Further, the demand response program used the data to explain variations in savings between equipment types and strategies. Although smart thermostats are becoming more popular and are being deployed in larger numbers than other connected devices, we estimate that other connected appliance and commercial energy management system data will offer similar benefits if evaluators can perfect these analytic techniques and if programs agree to embrace them.

Introduction – Just Another Demand Response Evaluation?

Smart thermostats are becoming more popular with consumers, in part because they offer the convenience of remote monitoring and control, and because they offer more user-friendly interfaces for programming. From an energy supply perspective, these devices can also offer demand response services to relieve pressure on the grid during periods of heavy load, typically during hot summer afternoons. During the summer seasons of 2015 and 2016, Lincoln Electric System and EPRI ran an innovative pilot that tested both smart thermostats and load control switches in a green-field setting.

The program's evaluation team devised a thermostat data analysis technique to use device-generated interval data in lieu of advanced metering infrastructure (AMI) data. We calculated the demand and energy savings from load cycling, temperature offset, and pre-cooling events in homes that used either traditional load control switches or communicating thermostats. This analysis helped inform the team about likely outcomes if the program were to follow different paths, scaling up from this pilot.

Hardware Smack-down: Smart Thermostats vs. Load Control Switches

The project team recruited just over 400 participants for the study (not counting approximately 30 employee participants who tested the technology during a beta phase). Participants could choose between smart thermostats and switches, until the stock of smart thermostats ran out; the remaining participants received only switches. The project team called 26 events over the summers of 2015 and 2016, not counting 5 test events. The team commanded the switches to cycle on and off during the event period of 3:00 pm to 6:00 pm. The team also tested different strategies on different days, with switches shutting down the HVAC compressors for between 33 percent of the time (cycling 20 minutes on, 10 minutes off) and 75 percent of the time (cycling 7.5 minutes on, 22.5 minutes off). The team controlled thermostats primarily through set-back events, either with or without a pre-cooling period from 1:00 pm to 3:00 pm, prior to the events. We called a few cycling events for the thermostats, to create a baseline for comparison between the two groups of homes. The impacts assessed included both average kW power reduction per home, as well as the increased load from pre-cooling and post-event snap-back. The team also calculated net energy impacts.

Belt and Suspenders: Quasi-experimental Design Validated with Experimental Design

The team used a quasi-experimental, within-subject design to calculate the final results of the study.[1] However, we called some events for only a random subset of participants, leaving the rest as a control group. For these events, we compared the impacts calculated through control-treatment comparison to the results generated by the within-subjects models, to validate the quasi-experimental method. This comparison showed that the quasi-experimental results, as seen in Figure 1, were typically a conservative assessment of the impacts for both switches and thermostats. We used them for reporting results on all events and for aggregate comparisons by event type. Although the results of the impact analysis and the details of the linear mixed-effects model were obviously central to the evaluation, they are not the focus of this paper.

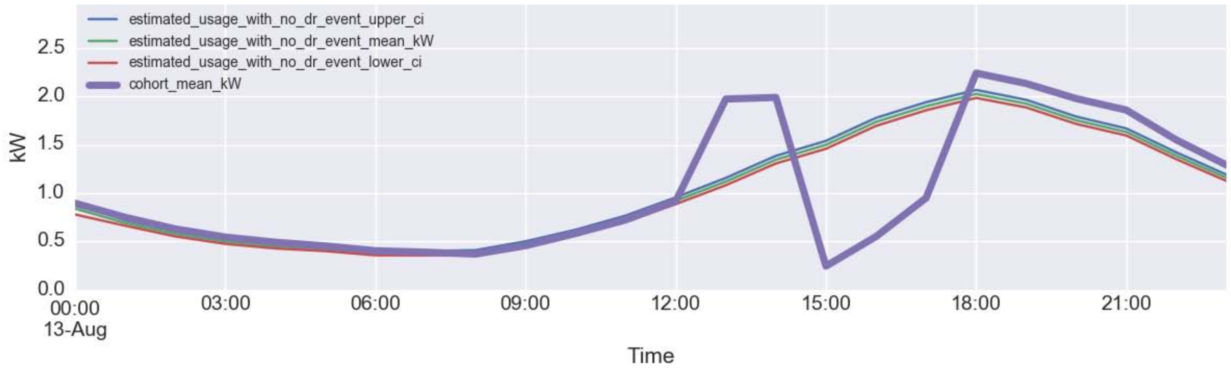


Figure 1. Pre-cooling event results, showing actual power trend (in purple), overlaid on modeled baseline with confidence interval.

The Bad News: No AMI Data. The Good News: Detailed Home Surveys, Including HVAC Nameplate Data

The utility conducting this pilot did not have AMI meters installed, and chose not to incur the considerable expense of installing interval power meters on participating homes during the pilot. However, the project team collected detailed survey data for each home and HVAC system, including their nameplate size, seasonal energy efficiency rating (SEER), and age. In addition, switches and thermostats both reported interval data. The switches recorded true power draw for the HVAC system (roughly) every seven minutes, but the thermostats recorded data only about their own operations, such as indoor temperature and calls for cooling and fan operation. The plan was to use the HVAC system ratings, outdoor air temperature from local weather stations, and the thermostats' run-time data to estimate

[1] Savings is calculated as the difference between each participant's event day and non-event day demand, adjusted for weather effects. See *State and Local Energy Efficiency Action Network, 2012* for more discussion.

power draw and event impacts. We validated and adjusted the approach, using amp loggers on a small sample of HVAC units.

Unknown Unknowns and the Analysts Who Love Them

As results began to emerge, the project team noted some surprises. This led to questions about whether the results were accurate, if they meant what we thought they did, and what this information indicated we should do to fix the analysis. We also asked what we had learned about the results of the pilot. Some examples:

- Online / offline status flag was not always reliably reported by the data collection system.
- Switch cycling was synchronizing HVAC loads, but randomizing start times resolved this issue.
- Time zone was different for device data, weather data and amp loggers, but was not labeled.
- Duty cycle could be used as an indicator for oversizing and might explain different impacts between different event strategies.

Note that we present the lessons learned as an example of the benefits and challenges of using device data, not as an indictment of either the technology or staff implementing this pilot. A flawless execution is surely everyone's goal and all preventable errors are regrettable; however, some are also inevitable. By using the available data to automatically analyze and visualize the result of each event, we believed it should be possible to detect such errors more quickly and diagnose them more reliably.

Methodology – Turning Data into Useful Results by Programming in Python

To parse and clean millions of switch and thermostat readings, convert raw thermostat data into power estimates, and model impacts and investigate anomalous results, we performed the evaluation using Python program code. This is the standard approach for interval data analytics at VEIC. We use not only the standard functions of Python 3.4, but also many of the libraries available with the Anaconda distribution, including Pandas for manipulating data-frames, Matplotlib for creating charts, and StatsModels for regression modeling. Initially we stored the data in a SQLite file-based database for ease of configuration. As the number of records grew with the second year's data, we moved the data to a Microsoft SQL Server database. We developed the code in Jupyter Notebooks, a web-based interactive development environment that allows the user to mix code, diagrams, and text / commentary in an executable "document." Automating the analysis process included exporting dozens of charts and tables for reporting and sharing with stakeholders.

Data Parsing and Cleaning

It is typically a straightforward process to parse data from xls, csv, xml, and txt format files; then to clean, reshape, and store the data in the database for later analysis. However, smart thermostat data (like that from many connected devices) are often collected as "event-based" data rather than as time-series data. So instead of rows containing the values of all variables at regular time intervals, the data consist of rows containing a timestamp (irregularly spaced or even duplicated for some rows), the variable name (for example, air temperature, cooling status) and the variable's value. Rows occur only when the value of the variable changes; for analog values, this typically is defined as a change of more than a certain minimum amount—for example, 0.5 degrees Fahrenheit for air temperature.

It would be extremely difficult in Excel to convert from this format to a typical time-series format, with rows at regular time intervals and columns for each variable, but it is straightforward in Python. The transformation occurs by assuming that numeric values are interpolated between readings and discrete values (like cooling on / off, or device online / offline status) are “forward-filled” or copied to all rows following an event record until the next event for that variable is recorded. Because the Python Pandas library contains functions that easily perform these operations, the transformation is not too challenging.

However, forward-filling is not an absolute rule, and analysts must consider the source of data and their meaning to determine how long it is safe to continue forward-filling values past the last reported updates. For example, is it reasonable to assume that the indoor temperature has not changed by even 0.5 degrees F, or that the air conditioning has been running non-stop for more than a day, simply because no new values are present for those variables during that time?

In the beginning of the data set from the smart thermostats, many devices went for weeks before reporting any values for “ResourceOnline,” the flag indicating if the thermostat was online and reporting data. Because it is a binary true / false field, we could have simply assumed that all timestamps before the first reported reading held the opposite value (that is, it must have been online before this time, if we are just now learning that it went offline). But we chose a more conservative approach of using the presence of indoor temperature readings, which are almost always reported to change at least once every few hours, to indicate if a device was online if no ResourceOnline value had yet been reported.

Also, we discovered that the central thermostat data collection system would go offline for 12 to 36 hours each month (but not at a fixed date or time) and would not update the ResourceOnline value at the beginning and end of this period to indicate the missing data. Consequently, the data conversion process assumed that the indoor temperature and HVAC system’s call for cooling status were unchanged for the entire period, since there were neither updated values nor a ResourceOnline flag of “false.”

These unmarked data gaps caused major inaccuracy in later analysis. Some HVAC units appeared to run continuously for 12 to 36 hours, whereas others were off for the entire time. Since the cooling run-time data from the thermostats were the basis for the “nameplate method” power calculations used for the impact estimates, these estimates would over-estimate the power trend during data outage periods. This could be seen clearly in data from devices that also calculated power trends from amperage data from dataloggers (Figure 2).

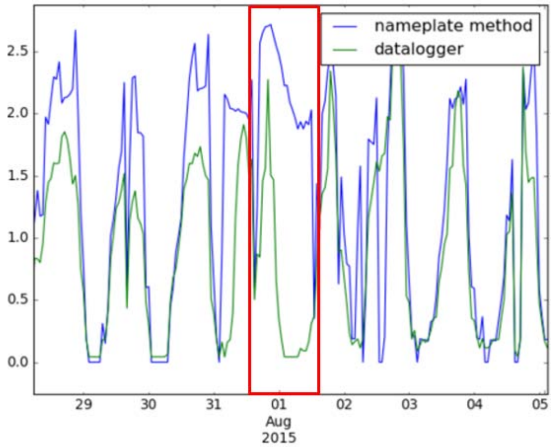


Figure 2. One week’s trend of power calculated from an amp logger compared to the nameplate method estimate.

As noted, the team used the indoor air temperature value (almost always updated every few hours) to calculate a proxy online status that could override the missing values. If 80 percent of devices did not report updated indoor air temperature values for at least six hours, we inferred that the data collection was down for maintenance. We marked the data for all devices during that period as “offline,” so that no power calculations were performed on data from those periods. A heat-map is shown in Figure

3, where the percent of devices with no indoor air temperature values reported in each hour appears in grayscale, and days proceed along the X-axis and hours from top to bottom on the Y-axis. The yellow bands show the outage periods detected by the "down for maintenance" algorithm. (Note the high percentage of non-updating temperature values in the pre-dawn hours; this is likely a normal feature of weather cycles and set-back schedules.)

One such maintenance period overlapped with a demand response event, rendering it impossible to calculate impacts for that event. For future pilots or programs, we recommend a contractual requirement of at least 99.9 percent uptime for a vendor's data collection system, and possibly an agreement that any scheduled maintenance take place outside critical performance periods such as demand response event windows.

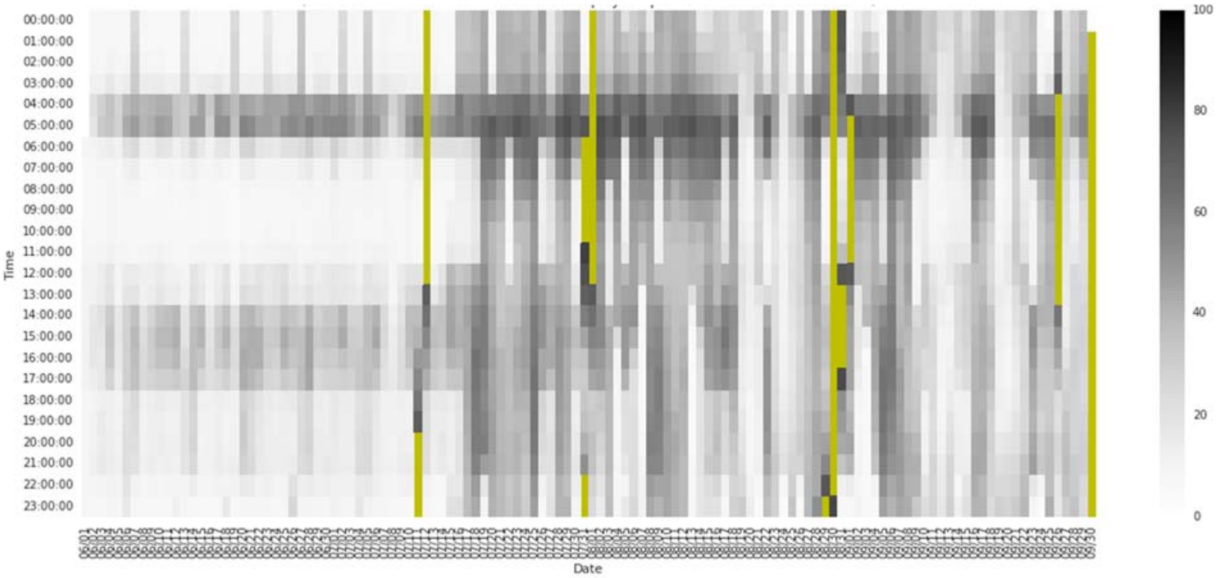


Figure 3. Percent of non-reporting devices; yellow bands indicate periods with more than 80% presumed outage.

Using HVAC Nameplate Data to Calculate Power

The underlying concept for this technique is familiar to anyone who has used HVAC system performance rating tables from datasheets. Each model's specifications involve the rated power draw at different outdoor air temperature conditions. Thus, for every HVAC device in the study, we could calculate the expected power draw for any given date and time, based on nameplate performance ratings and outdoor air temperature from local weather data, plus run-time data from the connected thermostat. This is referred to as the "nameplate method" for simplicity's sake.

In practice, missing and messy data required a few more steps before we could be sure that the calculated power trends were accurate enough for estimating impacts. Performance data are typically reported as the kW power draw at a range of temperature values spanning all operating conditions, but often spaced as far as every 10 degrees F. Rather than simply using the nearest value, we created a linear interpolation of each performance table and stored the resulting slope and intercept so that power could be more accurately estimated. These performance "curves" all appeared to be roughly linear (Figure 4), so other methods such as interpolation or polynomial fitting were not necessary. The team analyzed all units as though they were single stage, but the amp logger calibration should help to adjust for unknowns such as that. Since participants also reported age of units, we assumed that the efficiency of the units degraded by 1 percent per year. We incorporated this factor into the calculation for each HVAC unit.

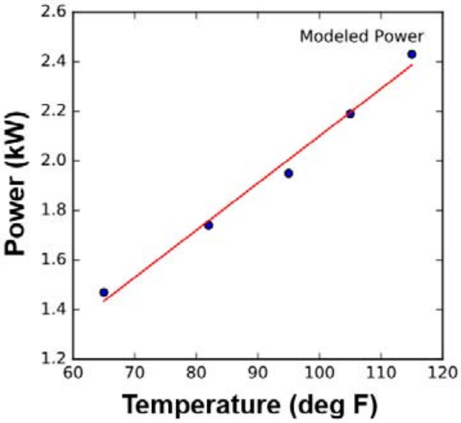


Figure 4. Linear performance curve, extrapolated from table in manufacturer's data sheet.

Filling missing values of size, age, or SEER

The straightforward example above also assumes the manufacturer's data sheet is available with performance tables for each HVAC unit in the study population. In fact, we found matching datasheets for just 12 units, whereas other units contained a size, SEER, and age (or even just a subset of those values). For the remaining units, the team found no datasheet for the model number, if any was even recorded on the survey. In those cases, we estimated values for size, SEER and age if any were missing. Then we looked up an average or interpolated performance curve for that size and SEER combination.

In analyzing the SEER and age for the population of HVAC units that had both values recorded, we noted there was a reliable relationship between the two that predictably showed that newer units have a higher SEER. This linear interpolation, shown in Figure 5, allowed the team to estimate either one of the SEER and age values, if the other was available. If both were missing, we used population averages. The HVAC unit size did not show a correlation with either SEER or age, or even with home size, so the population average size was assumed if size was missing. However, participants were most likely to document this parameter, and it was missing in fewer than 10 percent of cases.

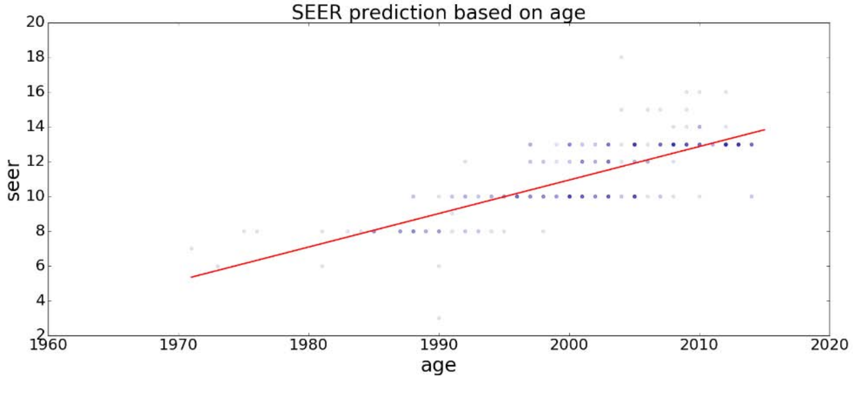


Figure 5. Linear approximation of SEER versus age for HVAC survey data in study sample.

We collected 42 HVAC datasheets and loaded the performance ratings into a database table. After ascertaining we had adequate coverage of data sheets across the range of size and SEER combinations in the pilot population, we calculated the slope and intercept for each datasheet's performance table. We then interpolated those values across the size / SEER plane to generate a performance curve for all participating HVAC units based on their size and SEER.

The performance curves from the HVAC data sheets represent the combined power of both the compressor and associated equipment in the "outdoor unit," and the power of the conditioned air circulation fan. In part of the analysis, we had to treat fan and compressor power separately. We subtracted the fan power from the estimate, assuming typical fan efficiency and capacity values of (*0.365W / CFM) * (400CFM / ton) * system tons* (Cutler 2013). This allowed for more accurate calculation of impacts in events where fans were commanded to run continuously while the compressor run-time was reduced.

Calibrating against interval metered amperage data

To validate and further refine the estimated power calculation from the nameplate method, we installed data loggers on approximately 40 HVAC units from the study population. Installed in two batches during the summers of 2015 and 2016, we collected data at 1-minute intervals for at least three weeks. The loggers monitored current draw in amps on the circuit feeding the "outdoor unit" of the HVAC system, including the compressor and associated fans and pumps, but not the conditioned air circulation fan. The amp measurements were used to calculate true power estimates from nominal voltage of 240V and an assumed power factor of 0.9, using the formula: *amps * 240V * 0.9 = power in watts*.

The team then compared these trends of amp logger-based compressor power to the compressor power trends calculated by the nameplate method. Visual inspection of the time-series trends revealed that most of the units showed very similar power levels from each method, although there were some exceptions—such as the unit shown in Figure 6, which appears to be a two-stage compressor rated according to the second-stage capacity.

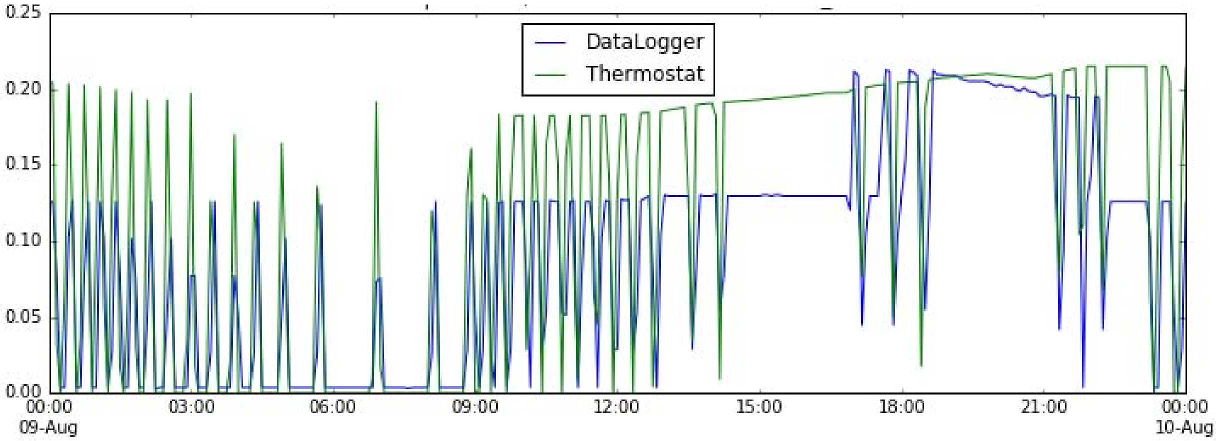


Figure 6. Comparison of nameplate method versus amp logger estimates for one day, poorly matching example.

When we plotted the hourly averages of amp logger-based power against the nameplate method-based power (*Thermostat* in Figure 7), the results suggested that there is a linear relationship between the power calculations from the two methods. If the amp logger power is the "ground truth" against which the nameplate method's results can be evaluated, we concluded that the results are capturing variability from outside air temperature and duty cycle; but we appeared to be miscalculating true power by some fixed factor. The "one-one line" on the plot represents equivalent values of nameplate and amp-logger-based power, but the regression line of the actual data has a slightly different slope, meaning that the nameplate method is overestimating power, compared to the amp logger calculation.

The average slope is 0.82 (a unitless metric) for the regression models from the HVAC units with amp logger data, but this sample shows a wide range of slopes. The histogram in Figure 8 shows the distribution, with 2015 and 2016 values presented separately. After we analyzed the data from 2015 (giving an average slope value of 0.71), we identified this adjustment factor value as a significant source of variability, so more loggers were deployed in 2016.

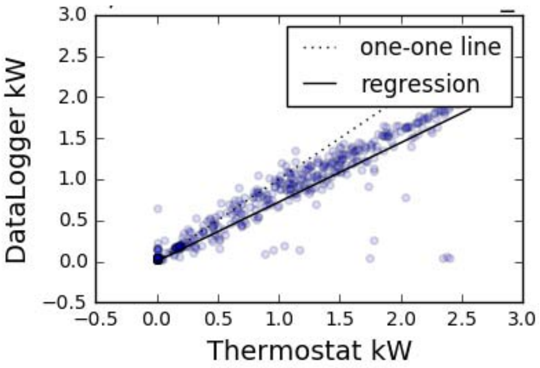


Figure 7. Comparison of hourly average power estimated by thermostat data and nameplate method, versus "ground truth" from amperage data logger.

The models for most units seem to provide a consistent ratio of actual to estimated power, with a regression model R^2 value that is almost always above 0.9. However, the variation in this slope between HVAC units still indicates that there is an unidentified source of variability in the nameplate method. We tested the relationship between the slope and the size, SEER, and age of the HVAC units, but none of those factors seems to be driving the variation in slope.

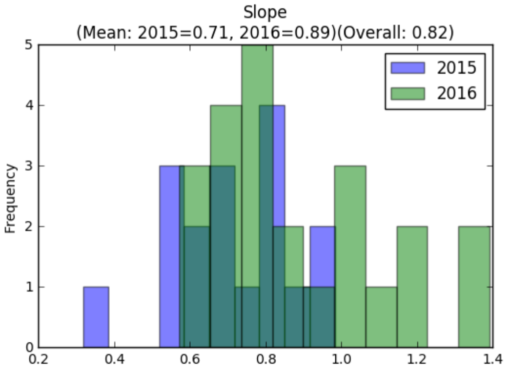


Figure 8. Histogram showing distribution of slope of regression model, relating nameplate method to logger-calculated power.

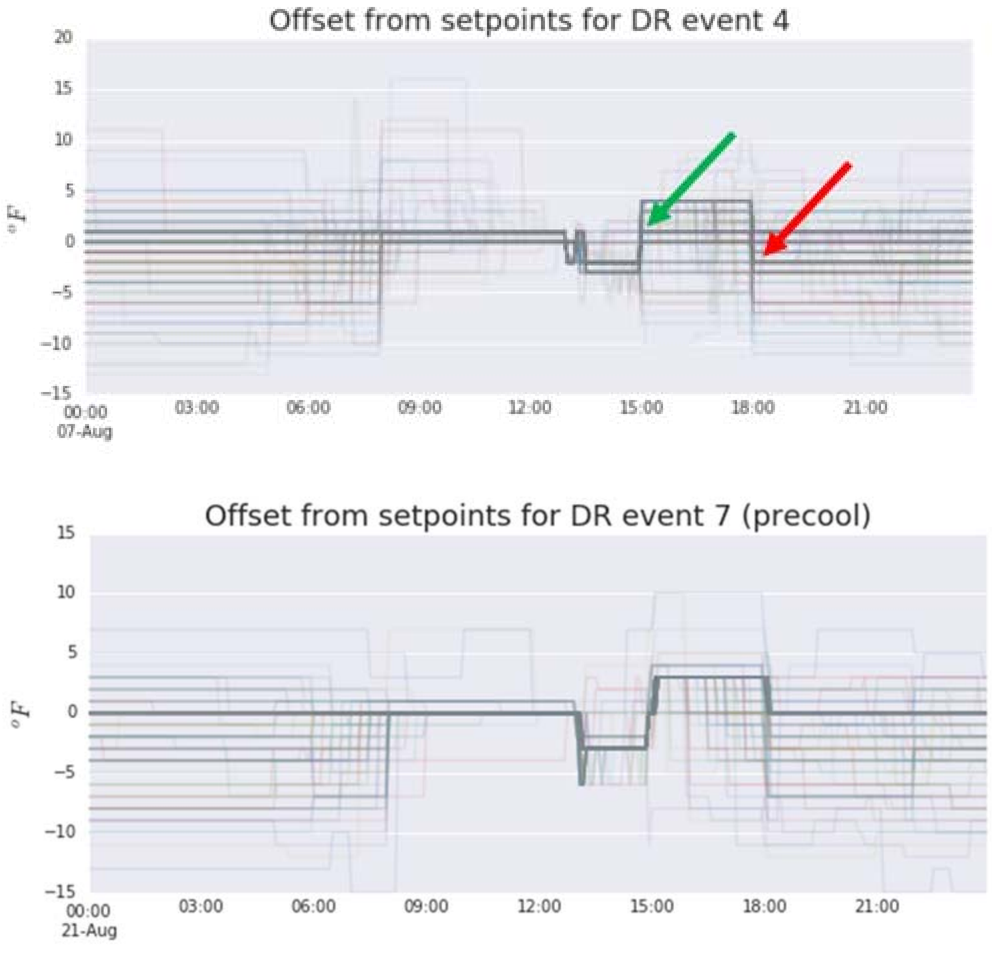
The Perils of Pre-cooling

One strategy tested for smart thermostats was "pre-cooling," in which the participating thermostats were set to cool the homes to below the customer-desired set-point before an event period, so the homes could coast for a longer time without added cooling before reaching the event set-point.

During standard offset events, set-points for each thermostat were each automatically adjusted to be 3 degrees F *warmer* than the customer had originally programmed for that day and time, then restored at the end of the event by dropping them back down by 3 degrees to the originally requested temperature. On pre-cool events, the set-point was adjusted to 3 degrees *cooler* than the programmed temperature, two hours before the event period. When the event period began, the temperature would be set to 3 degrees *warmer* than the programmed temperature, meaning a 6-degree increase in set-point at the 3:00 pm beginning of the event. An investigation of the thermostat data indicated that there might have been some confusion about how that rule was implemented in the first pre-cool event or two.

Some of the early pre-cool events showed increased energy consumption during pre-cooling but lower savings than expected, given the day's maximum outdoor temperature. A visualization of all participating devices' set-point trends, relative to the pre-event set-point (the day's set-point schedule was shifted to normalize the pre-event temperatures to zero, for all thermostats) revealed that the system apparently calculated the offsets incorrectly for some (but not all) of the devices. Further, it increased the set-point 3 degrees above the *pre-cool temperature,* rather than the *customer's intended temperature*, effectively enacting no setback for the event period (shown by the green arrow in Figure 9).

The post-event adjustment to lower all set-points by 3 degrees, which resulted in a lower temperature than the customer intended (shown by the red arrow in Figure 9), resulted in an increase in energy use relative to the baseline. Using the same visualization to examine the results for later events, this appeared to be a technical problem with the first event in particular, and was clearly resolved by the vendor in later events, as seen in the lower chart from Figure 9. As a side note, we marked these early events as *spoiled* and excluded them from any average impact calculations.


Offset from setpoints for DR event 4

Offset from setpoints for DR event 7 (precool)

Figure 9. Thermostat set-points relative to pre-event settings, showing demand response control strategies.

In later pre-cool events, we saw the correct execution of the strategy, visualized in the lower chart in Figure 9, as temperatures dropped 3 degrees for the pre-cool, increased by 6 degrees for the event, then dropped 3 degrees again to return to the original temperature (pegged to 0 degrees). Event 4 shows a brief deviation at the beginning of the pre-cool period, a known disruption to this event. However, the team could not determine whether the cause of or resolution to the disruption might be related to the problem. Another notable brief deviation in the other direction occurs at the beginning of the pre-cool period of Event 7; this was never explained but it did not seem to affect the event outcome.

Oversizing and Duty Cycle Analysis

The team observed some less-than-expected energy savings for switch devices, particularly on somewhat less hot days (the pilot included some extra test events on days with a maximum temperature below 90 degrees F). We traced them not to any issue with the execution of the pilot, but with the assumptions behind the cycling strategies.

Some switch events showed little or no savings, but clearly indicated that the correct participants received event signals and reacted appropriately to them. However, power levels initially shot up during the portion of the cycle in which the switch allowed the compressor to run, but seemed to subside before the "on" portion of the cycle was complete. This seemed somewhat odd. If the cycling strategy was going to produce savings, it should be cutting into the system run-time deeply enough that the compressors should be running for most or all of the "on" portions. If the HVAC systems were able to satisfy their set-points in less than the total on-time of each cycle, perhaps this 50 percent cycling strategy was not actually reducing run-time, but just shifting it somewhat.

To test this hypothesis, we calculated the baseline duty cycle for both switches and thermostats for each device as an hourly average. As anticipated, the average duty cycle for many of the devices was less than 50 percent, even on relatively hot, non-event days. Thus, a 50 percent cycling strategy would produce little or no savings for those homes, because they were sufficiently oversized to keep up with cooling demand, while running less than half the time, as shown in Figure 10 (upper plot).

Of course, the more aggressive 25 percent on / 75 percent off cycling strategy was more effective at producing savings. But such a cycling strategy risks more discomfort (and the associated opt-outs) for units that are not oversized. **By contrast, the smart thermostats' set-point temperature offset strategy could realize savings even on modestly hot days by reducing duty cycle for all devices without pushing comfort thresholds for any homes more than the others (Figure 10, lower plot).**

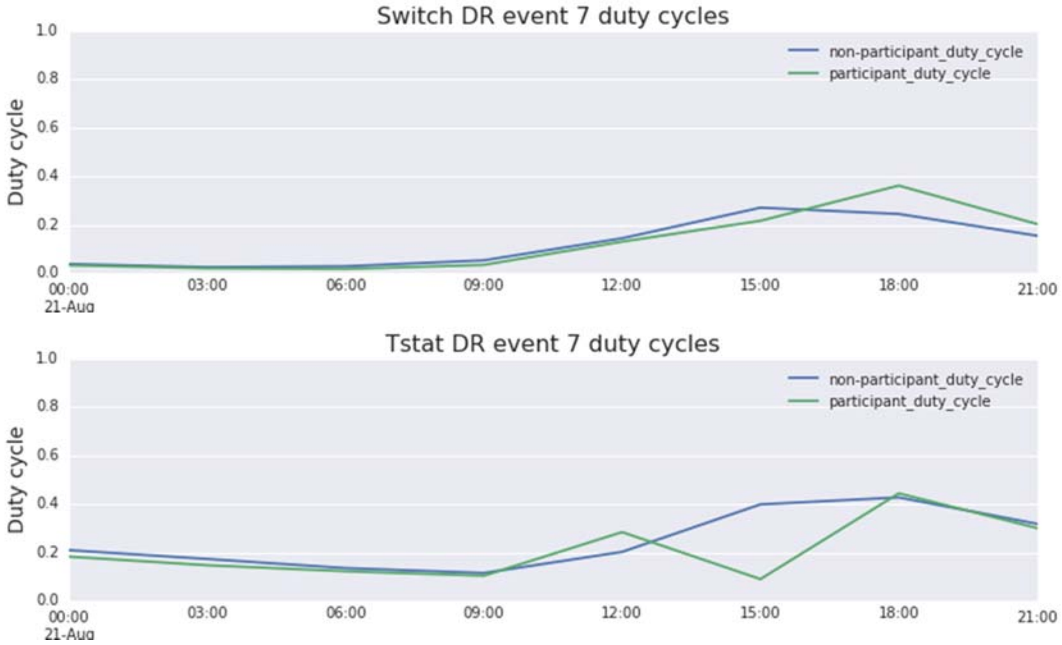


Figure 10. Event-day duty cycle trends for cycling and offset strategies under modest temperatures

Switch Cycling Synchronization

The team expected the smart thermostat data to be the focus of this evaluation, since the nameplate method would require significant post-processing. But the true power reported by switches could not simply be rolled up to hourly averages to report results. The switches reported data at a nominal 7-minute interval, but in practice the interval was not regular and typically registered a new value after less than 7 minutes if an event signal was transmitted, Sometimes they waited much longer than 7 minutes during periods of less activity (for example, overnight).

So the team converted the data from all switches to a regular 5-minute time series by copying reported readings to subsequent 5-minute intervals, up to a 15-minute limit for missing data. The 5-minute data set allowed a more detailed inspection of device behavior during events, revealing some unexpected behavior, as shown in Figure 11.

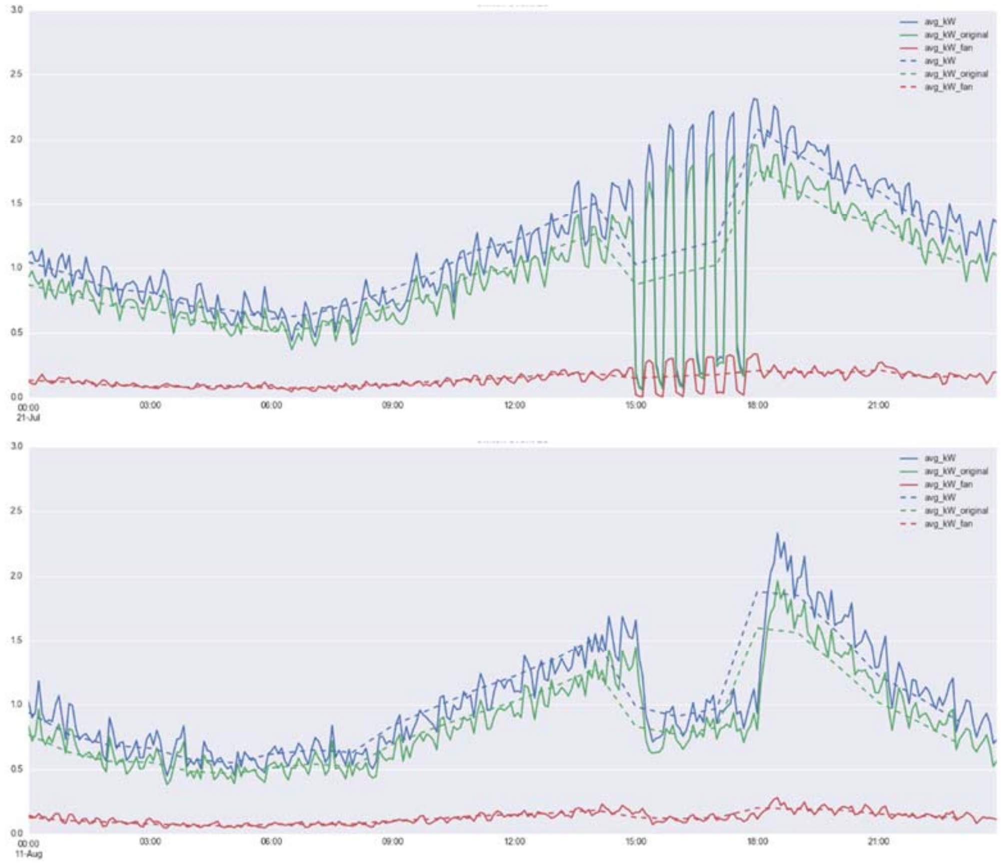


Figure 11. Diagnostic chart showing average power across all participating switch-controlled devices before start-time randomization (above) and after randomization (below). Total power (blue) is broken down by fan power (red) and compressor power (green), and is compared to hourly averages (dashed lines).

One of the first surprises in the 5-minute data was that switch cycling events were causing all 200 of the switch devices to turn on and off at the same exact time. These non-randomized-start events were causing sharp swings in demand (Figure 11, upper chart). The average hourly impact was still reducing demand on the grid, but the synchronized timing was causing large momentary spikes in demand that might have detrimental effects on grid stability, once the program scaled up.

The pilot team worked with the vendor to add a randomized start time to the events, so that different devices were being cycled at different times (Figure 11, lower chart). The impact of this change was clearly visible in the lack of sharp peaks within the event period. While the hourly average savings is similar for both events, at any given time during the event there is always a subset of units being cycled off, so the demand reduction is fairly consistent across the whole event period. Note that because thermostat-driven cycles are naturally staggered, it was not necessary to randomize the start time for thermostat setback events.

Conclusions

While this evaluation was obviously focused on the measurement of energy and demand impacts for the different technologies and strategies tested in the pilot, for the purposes of this paper we will simply say that the impact calculations were statistically significant, and well within the range of prior studies. Those quantitative results were certainly relevant to the utility program planners who used this pilot to plan future demand response investments, but there were several other lessons learned that might have even broader future applicability for the evaluation community.

First, the temperature offset strategies enabled by communicating thermostats are more effective at reducing demand across a wide range of conditions, particularly during more moderate events or when units are oversized. However, if using a cycling strategy, it is important to randomize start times to avoid adding volatility to a peak event. Further, care must be taken in designing, communicating, and executing strategies to avoid aberrant events like the pre-cool mishap described here. Various other similar issues that were uncovered during the course of this evaluation, such as the assignment of the group of accounts to receive an event, and the timing of on / off cycling are beyond the scope of this paper. Nevertheless, they reinforced the need for timely data-driven feedback to ensure the proper management of such distributed energy resources.

To calculate impacts and detect operational issues, the device-level data proved to be nearly as accurate as AMI data, and perhaps more valuable in some cases. The calibration of the nameplate method with amp loggers could certainly be improved with a larger data set and true power meters for comparison, even though this method might not be as accurate in situations where the nameplate ratings of each HVAC unit are not known. However, while 15-minute or even hourly AMI data might have detected the reduced impact of some events, the higher-frequency device data (including information about set-point, fan mode, event status, etc.) could shed better light on the root cause of the problem.

Parsing, storing, and analyzing these data through several layers of calculations and visualizations require different tools and skills than might be needed to perform a standard difference-in-differences analysis on power meter data. However, the many open-source software packages can make this approach far more accessible than it might have been a few years ago. Using a scripting language to set up this analysis will also allow us to generate results in real time, and help programs detect and correct issues as they arise—not just when the season has ended and the evaluation begins.

References

Cutler, D., J. Winkler, N. Kruis, C. Christensen, and M. Brandemuehl 2013. “Improved Modeling of Residential Air Conditioners and Heat Pumps for Energy Calculations.” NREL/TP-5500-56354, http://www.nrel.gov/docs/fy13osti/56354.pdf.

State and Local Energy Efficiency Action Network 2012. “Energy Efficiency Program Impact Evaluation Guide.” Prepared by Schiller Consulting, Inc. for SEE Action. http://www.seeaction.energy.gov.