“If the Shoe Fits”: Statistical Techniques When Traditional M&V Methods Don't Cut It

Rahi Shah, Warren Energy, New York, NY
Kevin Warren, Warren Energy, Lincoln University, PA
J.J. Costabile and Marcus Lewis, Con Edison, New York, NY

ABSTRACT

Measurement and Verification (M&V) is a crucial method for assessing performance of energy efficiency (EE) measures in large and complex facilities. Traditional M&V methods often rely on manual adjustments, linear models, and simplified assumptions to estimate baseline and reported period energy use, which can impact accuracy when analyzing complex systems and newer energy conservation technologies.

As building systems and technologies evolve, the nature of available data and the methods required to analyze it are also changing. Common sources of data for conventional M&V approaches are whole-facility metering, temporary data loggers, and onsite energy management systems each presenting their own challenges to data collection. While large-scale data logging can be economically unfeasible, EMS systems may not capture all the data required for analysis. Use of whole-facility billing data requires clean pre- and post-installation periods, relatively high savings, and only applies to measures with existing conditions baselines. The equipment installed in energy efficiency projects is increasingly connected to the web and includes onboard monitoring systems. “M&V 2.0” describes approaches including making use of new data-rich features of building systems.

The paper introduces refinements to conventional approaches that are facilitated by the increasing availability of performance data collected from web-enabled equipment (heat pumps and thermostats). We present case studies and comparative analyses for captured savings uncertainty for high dimensional data, categorized data for non-routine event detection, and hyperparameter tuning such as base temperatures compared to standard fixed degree day assumptions. New data streams can be integrated into savings analysis using machine-learning techniques. Additional options for accurate energy savings evaluations enable practitioners and policymakers better to assess EE projects in an increasingly complex energy landscape.

Introduction

The last decade has seen an explosion of new technologies and measures entering the EE sphere. Many of these incorporate data logging into their controls and this data is increasingly capable of being exported. Concurrently, some program managers have required M&V to verify claims, gather insights into measure mechanics and performance, and claim savings with a high level of certainty. M&V also remains an important component of Demand-Side Management program impact evaluation, particularly for custom programs. In response, M&V has evolved by leveraging new data sources and advances in widely accessible computational tools to enhance the scope, accuracy, and reliability of EE evaluations.

Utility program managers rely on M&V studies to help provide accurate energy savings for large and risky projects, as well as help evaluate new technologies that can expand and diversify program portfolios. When conducted as part of program implementation, accurate savings reporting from M&V studies help to improve impact evaluation realization rates. However, program managers often have competing interests with the resources that an M&V study requires, which can affect the accuracy of M&V results. For example, when multiple measures are installed at a site, the program may want savings for each measure but discerning the difference between those measures while doing an Option C billing analysis is not typically possible without additional metering. Program timelines may not allow for collection of a full year of post-installation data. Budgets may not allow for expensive data logger studies. Finding the correct calculation methodologies that serve the program team's limitations while also providing accurate results is one of M&V's most difficult challenges, and modifications to traditional M&V approaches are often required.

Option A and Option C, outlined by the International Performance Measurement and Verification Protocol (IPMVP), are among the most routinely applied approaches in energy efficiency M&V owing to their practicality, scalability, and alignment with common programmatic needs. Each offers evaluators specific advantages depending on project complexity, metering capabilities, and the level of analytical rigor required.

Option A involves measuring only the key parameters affected by the energy conservation measure (ECM), with other variables estimated based on specifications, historical data, or engineering judgment. Its simplicity and low cost make it attractive but also makes Option A less suitable for complex systems or ECMs where interactive effects and variable operational contexts can significantly affect performance.

Option C is routinely favored as it overcomes limitations of Option A by analyzing whole-building energy consumption using utility bills and/or submetering. This approach is particularly attractive in portfolio-level evaluations, performance contracting, or in programs targeting deep retrofits. The availability of tools that automate key tasks, such as weather normalization, further enhances its appeal. However, Option C also introduces methodological complexities. The accuracy of savings estimates depends on the robustness of baseline models, the resolution and completeness of available data, and the evaluator's ability to isolate the effects of ECMs from unrelated operational changes. Factors such as occupancy shifts or tenant behavior can introduce uncertainty and complicate attribution. Particularly, it is generally not applicable to new construction, change of building use, or major renovations, where no historical energy baseline exists. It is also unsuitable when existing systems were at or near end-of-life, as such cases typically require a replace-on-failure baseline using a counterfactual scenario based on code-compliant equipment.

The term "M&V 2.0" refers to emerging techniques that leverage high-resolution and nontraditional data sources, enabled by advances in web-connected technologies. Modern HVAC equipment now often includes onboard sensors and embedded data logging designed to provide real-time insights into system operation and occupancy. These capabilities can be leveraged for M&V. Much of the analysis in the case studies discussed below would have been cost-prohibitive or impossible a decade ago without these built-in capabilities. The advent of digital controls has also paved the way for more creative use of the available trend data for analyzing and isolating impact of energy conservation efforts in buildings.

This paper explores several machine learning methodologies designed to augment traditional IPMVP Option A and C analyses. It also presents case studies where conventional approaches proved insufficient, requiring a pivot to advanced data-driven techniques to achieve reliable results.

Addressing Limitations of Traditional Option C

IPMVP Option C traditionally uses monthly utility billing data, providing only 12 points per year. This low resolution often results in high fractional savings uncertainty (FSU), especially for buildings with variable energy use (ASHRAE Guideline 14-2014). Programs also rarely delay incentive payments for a full year, making reliance on monthly data impractical for program implementers. If sufficient weather variation is captured within six months, daily or hourly data can be used, though these increase noise and regression bias when baseline and post periods are weather-skewed.

Establishing a clean baseline is another challenge. Accurate regression requires a period with full weather variability but no operational disruptions. In large facilities, overlapping ECMs, failures, or atypical events often distort baselines. The authors have encountered this situation many times with multifamily complexes that installed multiple projects or had major system faults. This reduces reliability and complicates attribution of savings to specific ECMs, since Option C captures only whole-building use. Occupancy changes, lighting or plug load shifts, and other non-related variations further confound results.

Weather normalization also adds complexity. M&V practitioners commonly use base-65 degree days because they are readily available, but fixed 65°F degree days often fit poorly. Variable-base degree days and change-point models improve accuracy. Linear models, even with site-specific base temperatures will fail to capture nonlinear system behavior, such as heat pumps with resistance backup. More advanced segmented or ensemble models may be required.

Ultimately, while Option C is a powerful technique, it is often limited by data constraints, baseline uncertainty, and end-use attribution challenges. Successful application requires early stakeholder engagement, careful identification of non-routine events, and flexible, data-driven modeling approaches beyond traditional regressions.

The following sections introduce techniques and examples where additional data sources and analysis techniques were used to overcome limitations or enhance our understanding of the impact of measure implemented.

Machine Learning

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on developing algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed for every specific task (Samuel 1959; Mitchell 1997). However, while traditional statistics and M&V use the modeling tool for inference, ML typically focuses on the predictive performance of the tool on new data.

A growing body of literature has shown that empirical modeling could outperform engineering-based modeling when data preprocessing, model development, learning algorithm and post-processing are executed properly (Im et al. 2020).

A typical ML approach includes the following steps:

- Data Collection and Preprocessing: Gather raw data from relevant sources, then clean, align, and format it for analysis. This includes handling missing values, correcting errors, and standardizing time intervals or units. For M&V studies, some of the common sources of additional data sources can be building management systems (BMS) and vendor installed sub-metering.
- Feature Engineering & Selection: Create meaningful input variables (features) that capture key patterns in the data. Select only the most relevant features to improve model accuracy and avoid overfitting.
- Model Training: Use historical data to train a model that learns the relationship between inputs and outputs. The model adjusts its parameters to minimize prediction error and generalize to new data.

Feature engineering and selection, which are discussed in more detail later, are similar to the traditional M&V baseline and post-installation data preparation process but offer greater flexibility by allowing the use of a broader range of data sources and transformations. Similarly, model training mirrors the development of baseline and post-installation models described in the previous section. Thanks to the rapid growth of open-source tools over the past five years, implementing even complex algorithms has become more accessible, enabling experimentation and exploration across a wide range of skill levels.

Feature engineering is the process of creating, transforming, or selecting the input variables, known as features, that improve the performance of a model. This can include creating new features which capture site energy consumption patterns that can often better explain model performance. For example, replacing degree days by hourly occupancy rate X degree days for full building energy usage or using hourly compressor run hours X equipment cooling capacity to determine building cooling loads might provide a better model rather than using total electrical consumption for the site.

A common use case of feature engineering is using categorical variables such as unit status and weekday/weekends into the model through a process called encoding. Encoding transforms categorical data, which are non-numeric by nature, into numerical representations. Common techniques include one-hot encoding, which creates binary columns for each category, and label encoding, which assigns a unique integer to each category. This enables models to leverage valuable qualitative information, improving their ability to capture patterns related to operational status or temporal effects.

Finally, selecting the most relevant features that improve model performance without adding unnecessary complexity is essential when using multiple variables. It often involves dropping or shrinking variables whose predictive value is limited. Common methods for feature selection are discussed further on. However, it is necessary to take the primary objective of the study and available data points into consideration before applying these strategies. Without context, increasing level of model complexity can run into issues of over fitting, lack of interpretation and increase of bias. In the following section, we see examples of these concepts applied in various projects we have encountered in our studies.

Base Temperature Selection

Some of the most commonly used independent variables used in M&V studies are heating and cooling degree days. To calculate them, one needs to determine a “base temperature”. An accepted default value is often pre-determined based on the region by program managers within a state authority or utility for all sites within their jurisdiction. For instance, the base temperature is 65°F in the Michigan TRM, whereas it is 62°F for heating and 57.5F for cooling as defined by Con Edison in their New York territory. Specifically for Con Edison, base temperatures were selected to ensure consistency with corporate reporting, such as forecasting, and are referenced in internal procedures as well as in Appendix 2 of Con Edison’s Climate Change Vulnerability Study. These reference temperatures differ from the base 65°F assumption outlined in the New York State Technical Resource Manual and in NYSERDA’s Monthly Colling and Heating Degree Day Data.

The base temperature, also called the balance point temperature, represents the outdoor temperature below which heating is needed or above which cooling is required. Since the temperatures define an equilibrium between heat loss and heat gain for a building, they vary between neighboring sites, given that building envelope, insulation and orientation can vary. The assumed or pre-determined method is practiced prevalently, but it does not reflect the specific characteristics of a building.

When site-specific savings are needed, such as for projects with large savings, a more rigorous approach is to determine the base temperature empirically, i.e. testing a range of possible base temperatures, and calculating degree days for each one. During a typical Option C analysis, we are effectively building a simple linear regression model for a given site’s energy consumption for a particular period. The goal is a robust model which maximizes accuracy and minimizes the error function. At times, site data such as information regarding a site’s HVAC system and operation can provide insights to improve model design, namely as piece-wise linear regression or non-linear models.

Linear Model Design

Billing analyses may reveal nonlinear patterns that a single regression curve over the entire range of the independent variable could miss. A refinement to address this is to introduce multiple switchover temperatures and develop piecewise or aggregated linear models, i.e., separate linear models for different temperature bands.

For example, HVAC systems may use different fuels or heating/cooling mechanisms for base and peak conditions, leading to distinct energy-use behaviors under different outdoor air temperatures

(OATs). A single linear model forced to fit the entire range may underfit in some regions or overfit in others. By segmenting the data (e.g., by OAT bins) and fitting individual models to each segment, each model can more accurately capture the local trend within its range. Aggregating these models results in overall predictions that better reflect actual system behavior across varying conditions, thereby reducing bias and improving both model fit and forecasting accuracy.

Model performance can be evaluated using metrics such as the coefficient of determination (R^2), fractional savings uncertainty (FSU), coefficient of variation of the root mean square error (CV-RMSE), and other relevant error statistics.

Below are some examples of Option C models where using the two methods discussed earlier (custom site-specific balance temperatures and piecewise regressions) led to significant improvements in model performance. Figure 1 shows gas use in a campus with a single gas-fired boiler plant serving steam radiators across the site. There are clearly two distinct operating modes. Use of piecewise regressions leads to improved model performance and tracking of actual consumption when compared to use of a simple linear regression.

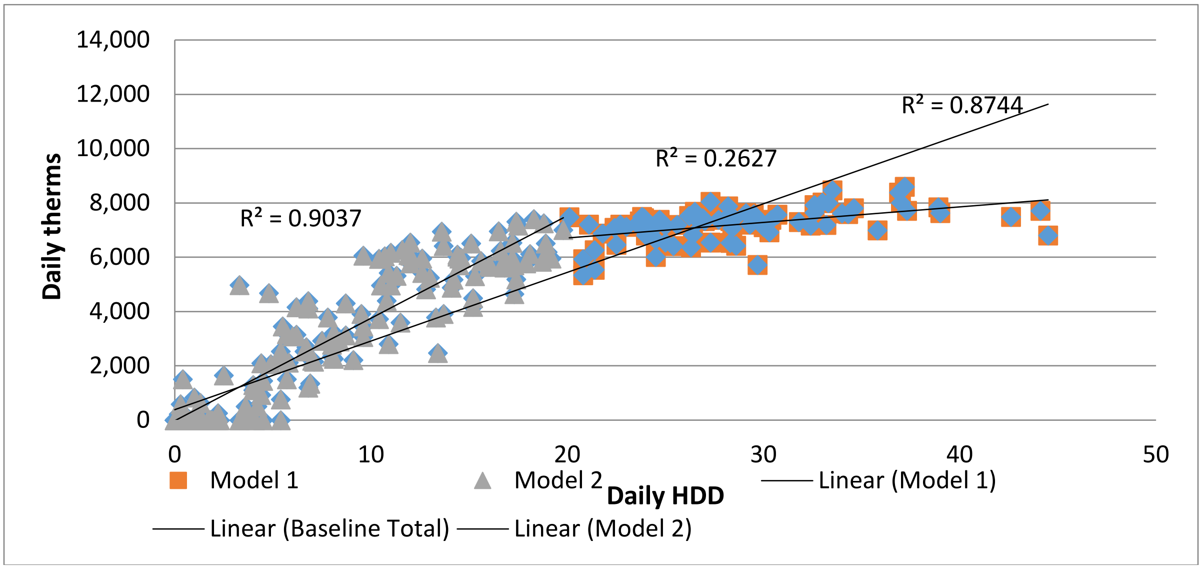

Figure 1. Comparing effect of model design on predictive performance for Site T from Table 1.

The R^2 for the simple linear regression is 0.87. For the two piecewise regressions, one has a higher R^2 and one a lower. If we were to use R^2 as a metric to decide the fitness, it is unclear which technique is superior; however, if we consider CVRMSE and FSU, the superiority of the piecewise approach is clear. As we can see from Table 1 below, an improvement in R^2 is not a great indicator for an improvement in overall uncertainty.

Table 1 shows the impact of using custom site-specific balance temperatures vs use of the 65F standard for ten multifamily sites for which Option C analysis was conducted. Each of these involved the analysis of hourly AMI gas data resolved to daily intervals. In every case, the FSU met (or nearly met) the requirements of ASHRAE Guideline 14 with FSU below 50% even with the use of the standard baseline[1]. However, for all of the facilities, the lowest FSU was obtained with a balance temperature other than 65. In most cases, the improvement in savings precision was greater than 15%.

[1] In the author's experience, NYC multifamily properties have more predictable consumption profiles than other types of facilities, so the low FSU value should not be taken as typical for other facility types. This is usually the result of boiler controllers programmed with sequences that provide operating times that are linear with outside air.

Table 1. Selection of multifamily projects with Option C analysis in New York City, with and without base temperature selection

Site	Pre-determined Base Temperature			Base Temperature Optimization			Change in Base Temperature (F)	Improvement (%)		
	R^2	CV (RMSE)	FSU	R^2	CV (RMSE)	FSU		R^2	CV (RMSE)	FSU
P	0.99	5.6%	28.3%	0.99	2.9%	14.7%	3	0%	48%	48%
Q	0.80	28.7%	26.9%	0.82	27.2%	19.1%	5	2%	5%	29%
R	0.95	7.3%	9.9%	0.95	5.9%	8.0%	3	0%	19%	19%
S	0.93	15.0%	53.4%	0.95	13.1%	42.6%	7	2%	13%	20%
T	0.95	25.9%	12.8%	0.95	25.7%	12.7%	2	0%	1%	1%
U	0.87	16.1%	44.0%	0.87	15.8%	43.7%	7	0%	2%	1%
V	0.91	14.6%	6.4%	0.93	13.8%	6.1%	7	2%	5%	5%
W	0.97	14.3%	29.2%	0.98	13.3%	27.3%	6	1%	7%	7%
X	0.96	13.0%	49.0%	0.98	8.8%	33.0%	5	2%	32%	33%
Y	0.95	18.8%	48.0%	0.97	15.2%	38.7%	6	2%	19%	19%

Use of optimal site-specific base temperatures ensures the energy model reflects how the building truly responds to the weather, which leads to more reliable savings estimates. Even though not universally practiced, this process is fairly easy to execute as base temperatures typically fall between 55- and 75-degrees F. However, it can become more complicated once additional independent variables are used in the model which brings us to our next topic.

Independent Variable/Feature Selection

For any energy efficiency analysis, one can list many features available to evaluate and explain energy model performance. Some examples include loop temperatures, degree days, fan/pump speed, occupancy rate, etc., but additional parameters do not always improve model performance and can often lead to other issues such as lack of model interpretability and overfitting to a certain set of data. For most M&V studies, model interpretability can be a powerful tool to review new and upcoming technologies.

Feature selection is as important as it can be complex and depending on the type of available data (time series data versus static observations, stationary versus seasonal data) and model construction (linear regression, logistic regression, decision trees, support vector machines, etc.), one can select an appropriate algorithm from the following categories:

- Filter Methods: Features which have least correlation to dependent data can be dropped. Examples: Analysis of Variance ANOVA (F-test), Chi-square tests, etc.
- Wrapper Methods: use a model to evaluate subsets and drop features that have the smallest coefficients. Example: Recursive feature elimination.
- Embedded Methods/Shrinkage Methods: selection happens during model training, i.e., as the model is fit to the data, features are ranked and their coefficients are sized according to the importance of each feature. Examples include regularization approaches such as lasso and ridge regression or tree-based selection methods such as gradient boosted methods.
- Dimension reduction methods such as principal component analysis.

Detection of Anomalies and Non-Routine Events

Non routine events (NREs)[2] are commonly known to obfuscate analysis and if not addressed will inflate or deflate savings estimates. Some of the most common ways for detecting anomalies are visually, by reviewing time-series energy data for each period, site inspections, and interviews. Common examples include lack of utility readings, recording errors, leaks, and more. Anomalies and outliers lead to higher standard error which can directly affect confidence error and FSU values. However, as model complexity increases, two dimensional plots can become more complicated. Visual methods of detection may fail. In an analysis including multiple variables, each variable could lie within an acceptable range, but with a highly unlikely coincidence value, e.g., a data point with a high peak temperature and high heating degree day. We suggest the following two techniques to detect outliers as well as high leverage points:

- Studentized residual: We typically expect studentized residual values to range between -3 and +3 in well-behaved regression models, as values outside this range may indicate outliers or points with high influence (James et al. 2021).
- Leverage statistic: Determines distance of an observation from mean observations.

Plotting leverage statistic and studentized residuals against each other can be a powerful tool for anomaly detection in linear (and some non-linear) models. Other, model independent methods include Component analysis, k-Nearest neighbor, etc.

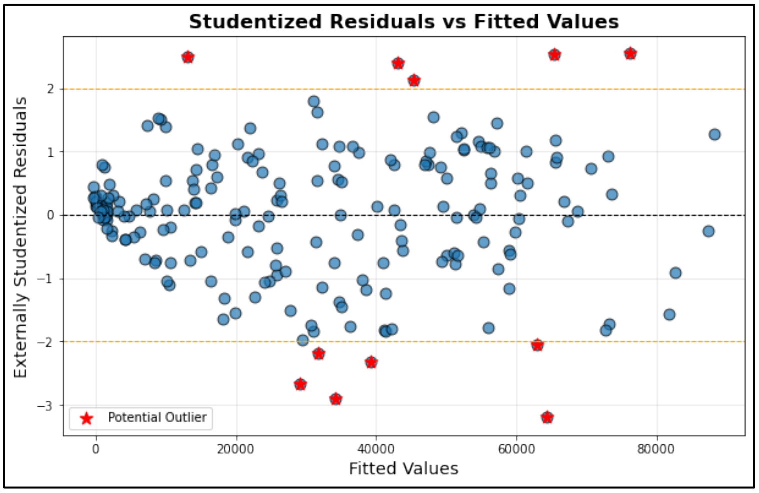

Figure 2. Sample plot of studentized residuals versus fitted values where independent variables such as HDD and solar irradiation are used to predict total heating load on the site.

Scale Normalization

Feature engineering and addition of new data may lead to situations where multiple features have a statistically significant impact on the energy model such as degree days, outdoor air temperature, occupancy (%), setpoints, solar irradiance, etc. Feature normalization is an important step in processing raw trend data for multivariable regression, especially when the features have different units, scales, or

[2] Non routine events are defined as "A change in static factors that is not part of the program and that affects energy use in the facility, such as a change in building occupancy, hours of use, equipment, or major process change." (IPMVP, EVO 2022, Volume I, Section 3.6.3)

ranges. Without normalization, variables with larger magnitudes can disproportionately influence the model, leading to biased or unstable regression results (James et al. 2021).

Unless all the features share a similar scale (i.e., mean and standard deviation), features with larger averages or deviation will tend to dominate when directly fed into a linear regression or feature selection algorithm. One of the most common approaches to normalizing data is using Z-score normalization. The algorithm for Z-score normalization is as follows:

$$x' = \frac{(x - \mu)}{\sigma}$$

Where:

- μ: mean of the feature
- σ: standard deviation of the feature

Algorithms like these can be executed in Excel when the number of features is small and well understood. However, there are scenarios where the quantity and complexity of data necessitate advanced tools such as Python, refer to Site A data in the case studies section below where more than 2 million data points were used to create a regression model.

In addition to the ones noted above, there are several useful techniques available while working with time-series data sets as commonly seen in M&V projects such as:

- Using nonlinear models, especially when multiple variables are involved, can often lead to noticeable improvements in performance. Some common examples include decision trees like random forests and gradient boosted trees, as well as piecewise polynomial models and support vector machines. That said, it is important to understand the trade-offs. These more advanced models can be harder to interpret, may be more prone to overfitting, and might not work as well with standard weather normalization methods.
- Resampling methods such as cross-validation can be used with static and time series data where model performance is tested on subsets of baseline or post data instead of the whole data set to check for overfitting and model comparison and selection.
- Autocorrelation is important to test data in high-frequency time series (example: Durbin-Watson), as it reflects dependence between values or errors across timestamps. If unaddressed, it can bias model estimates and reduce predictive accuracy. Techniques like differencing or including lag terms help mitigate its impact.
- Anomaly detection tests such as Z-score and seasonal decomposition for univariate time series datasets, Hotelling's distance for multivariate models, support vectors machines and principal component analysis for non-linear high dimensional data sets.

Case Studies

Next, we present case studies that highlight real-world challenges we've encountered and demonstrate how these concepts were applied to address them effectively. Each site leveraged the embedded data collection capabilities of the installed equipment whether VRF system controls, smart thermostats or chiller controls to obtain detailed performance data that, until recently, would have required a large-scale and costly metering effort. Each of these M&V studies were conducted within the program implementation and formed the basis for the payment of large custom incentives.

Site A: New Construction Clean Heat. Sites A is new construction multi-family residential development in the Bronx, NY, consisting of several high-rise towers with shared amenities such as gyms, game rooms, and lounges. Air-source VRF heat pumps were installed rather than a conventional fossil fuel heating system, making the project eligible for “clean heat” incentives. We performed a detailed evaluation of energy savings relative to a code-minimum baseline. Heat Pumps were compared to a baseline HVAC configuration obtained from ASHRAE Appendix G.

A traditional Option C was not a feasible option due to the site being a new development. An Excel-based tool developed by the Con Edison Clean Heat program was used to create temperature bin-based baseline and proposed operating profiles, resulting in large projected gas savings. Given the size of the potential incentive, the program team requested a more in-depth M&V study to limit uncertainty.

The utility was able to provide 15-minute interval AMI electric metering data for each individual account present onsite where the heat pump compressor and condenser fan were connected to a handful of meters while the evaporator fans were fed by each tenant meters. The AMI data alone might have been sufficient to develop profiles of post-installation electricity use of the VRF system, but we would not have known the proportion of energy responsible for heating and cooling related consumption while the changing site occupancy would have further obfuscated the weather dependent regression models. Development of the heating and cooling load profiles, necessary for calculating the baseline energy use, would have required the use of several assumptions. Thankfully, an initial inspection found that the VRF system’s compressor and condenser energy consumption were trended for each apartment unit in hourly intervals for occupied as well as unoccupied units for billing purposes. This was an embedded feature of the chosen heat pumps that allowed for an M&V 2.0 approach.

This allowed the reviewer to make full use of eight months of submetered data from about 490 individual spaces, adding up to roughly 2.8 million data points. The goal was to create detailed heating and cooling energy profiles. When such a large dataset is used for analysis, however, data cleaning and preprocessing becomes a vital step before modeling is performed. The following techniques were performed to aid in post-installation data processing and provide more accurate results:

- The site was unable to provide detailed occupancy data for each of their units but rather provided monthly occupancy rates across the site. Thus, the hourly VRF data from each apartment/common space was evaluated to confirm continuous occupation and to filter out units which were likely un-leased for any portion of time. If the room did not meet a minimum cut-off requirement for consumption in a week/month, the room was dropped from analysis for the given period.
- Outliers and high leverage points were detected and removed from the sample.
- The occupancy rate derived from metering data was married to the site occupancy data as an additional dependent variable in addition to typical weather variables. The program team determined that incentives needed to be based on end-of-the year occupancy rates. As noted earlier, the variables being in different scales required scale normalization.
- Equipment efficiency curves from the manufacturer and weather data was used to project current and future building cooling and heating loads.
- These loads were then normalized and used to calibrate the Option A tool for clean heat projects at end of the year occupancy rate.

The approach allowed us to not only capture first year savings but also predict, with a high degree of certainty, the projected savings at a future date when the site would reach full occupancy. The correction in Option A for site A approaches following this analysis was as large as 28.4% as compared to traditional engineering based approach.[3] In addition to this, we also understood that the program team assumption for cooling load to dominate at the site was corrected in newer revisions of the Option A tool.

[3] The correction (%) rates are based on overall MMBtu difference between pre-install analysis based on Option A and final reported savings estimates.

Thus, the use of embedded data collection features were exploited to overcome challenges related to conventional methods.

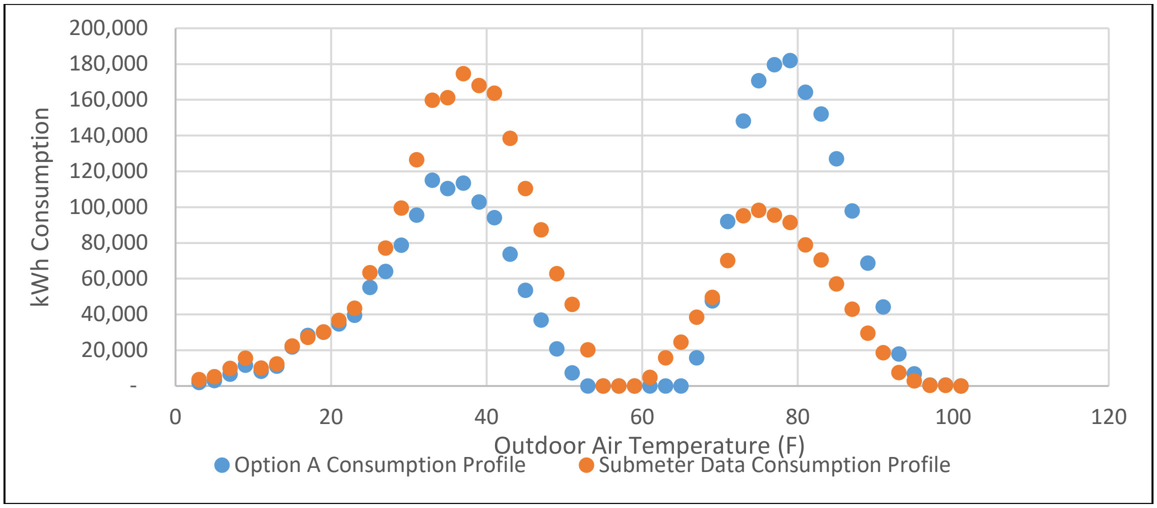

Figure 3. Comparing Site A's energy consumption patterns as predicted by Con Edison's Clean Heat Tool versus the site's submetered data.

Site B: Hotel in Midtown Manhattan. Site B is a 518-room, high-end, 25-story, high-rise hotel in midtown Manhattan. The site utilized water source heat pumps (WSHPs) to provide heating or cooling in the guest rooms with local thermostats without any occupancy sensors. This led to a waste of energy when the guests and staff often left units operating around the clock, even as the rooms were unoccupied. The primary ECM at the site was the installation of EMS[4]-integrated thermostats for existing heat pumps in rooms that would also have occupancy sensors installed and be integrated with the PMS[5] (Property Management System). This combination allowed for the ability to increase setback temperatures for non-rented and unoccupied rooms. The setback temperatures were calculated by the EMS system such that the recovery time for each room, once occupancy was established, would be below 20 minutes.

The hotel had been closed due to construction work and had been reopened for only a couple of months before the project was selected for M&V. There was no recent period with significant occupancy that could be used as a baseline period. Also, when the project was selected, the occupancy rate was 85% and climbing which presented challenges for characterizing the post-installation performance. The limited data during the baseline period made Option C unfeasible. Fortunately, the contractor informed us that proposed smart thermostats had trending and bypass functionality.

This led to a unique opportunity to conduct in-situ M&V analysis by conducting a random experiment. We randomly selected a sample of rooms to operate as baseline and another sample to operate as proposed. With a large enough sample size to population ratio, the project savings could be determined with a known level of uncertainty. A sample of 4 floors of rooms in bypass (baseline) mode and 4 in operational mode were selected. Data regarding room occupancy, fan and compressor run hours as well as room setpoints were collected from the thermostats. This data, along with unit specifications were used to create an average room energy consumption profile which could then be extrapolated to the rest of the site.

During unoccupied periods, the WSHPs in bypassed rooms were expected to operate more than the WSHPs in rooms in the post-ECM sample. This was observed in the data. However, the operation of

[4] Energy Management System.

[5] The property management system is used for scheduling suits and guest management.

the two samples also differed during occupied periods. This was not expected; both samples were expected to operate similarly during an occupied scenario. Instead, the collected data showed that the units in the baseline sample ran much more when the room was designated as occupied compared to the post-ECM sample. A simple comparison of baseline and performance data indicated savings, but no savings could be reasonably attributed to the controls measure for these conditions. An alternative to the M&V 2.0 strategy was to install amp or kW loggers on a significant sample of units in the baseline and performance floors. This data would have shown savings, but without the thermostat data, we would have been unaware that some of the "savings" was occurring during occupied periods (the occupied period for a room is not a fixed set of hours and could occur in any hour). The savings from the alternative method would have been overstated.

Through staff interviews, it was determined that this discrepancy was likely due to housekeeping staff manually adjusting unit setpoints. Prior to the retrofit, housekeeping staff were told to manually set back thermostats. Once the ECM was installed, they were no longer required to do this. Unfortunately, the bypassed units did not perfectly represent the pre-installation conditions, so it appears that the heat pumps were interacted with by both the hotel guests and employees. A baseline adjustment was used to eliminate this effect. The high-resolution data from the thermostats allowed for quantification of the adjustment. Without the data collection feature of the thermostats, savings would have been overstated.

Depending on the time of the day, a sample of rooms could be in cooling mode simultaneously with others in heating mode leading to interactions between the heat pumps and condenser water system. Collection of room occupancy, equipment mode and run hours allowed for creation of an average load profile for the site. The data allowed for determination of the heating and cooling profiles separately so the overlap in the profiles could be observed. Our analysis allowed us to create a unique consumption profile for the site that captured the various interactive elements and determined that the overall the impact of smart thermostats exceeded initial engineering estimates significantly. The final correction in savings were reported 102% more than reported by the applicant and ~350% more than previously calculated through a conservative engineering approach. Figures below depict the difference in heating and cooling load curves as determined through sampling.

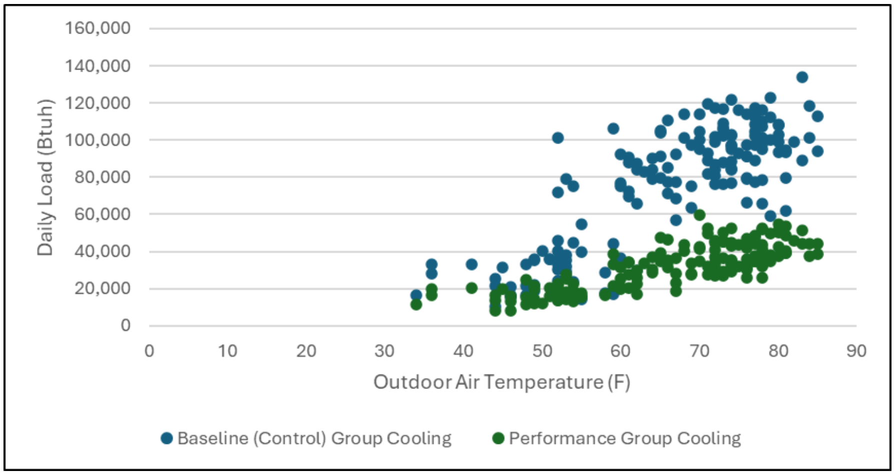

Figure 4a. Calculated difference in heating loads between control and performance groups.

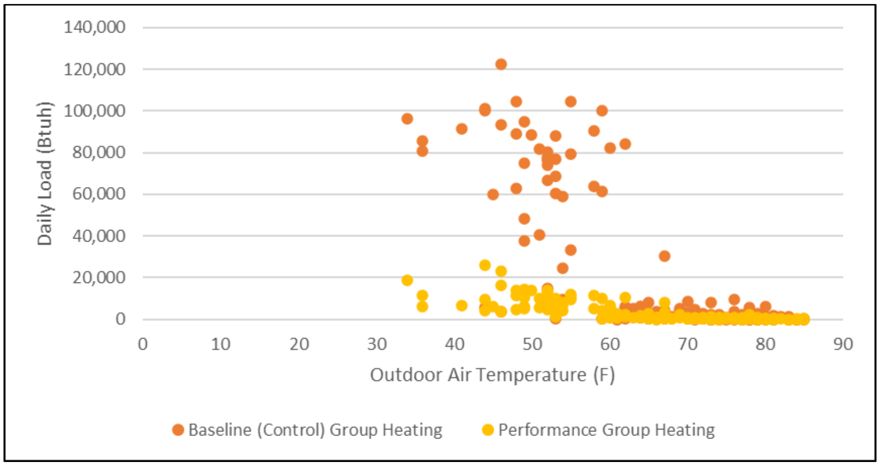

Figure 4b. Calculated difference in cooling loads between control and performance groups.

The case study is another example where feature engineering on data collected from embedded systems was utilized to increase accuracy and overcome challenges posed by conventional engineering or M&V methods.

Site C: Commercial high-rise in midtown Manhattan. A 27-story commercial high-rise in Manhattan, NY, was selected by the Con Edison Pilots Program to evaluate the savings potential of a new fluid treatment system, projected to improve chiller operating efficiency by approximately 15%. However, the site lacked both a functional building automation system (BAS) and detailed performance ratings of the baseline system. In addition, because the mechanical plant was not submetered and site occupancy patterns varied, the relatively modest savings potential (estimated at less than 10%) made the project a poor candidate for Option C analysis.

With some convincing, the project owners agreed to initially operate the proposed system in bypass mode to allow for trend data collection to collect baseline chiller data for a period of 3 months before switching over to the performance period. This also required calibration of the existing BAS so that data points such as chiller amps, BTU meters, entering and leaving temperatures on both evaporator and condenser water side could be collected during this period to capture chilled water system operation.

It is common to use bi-quadratic curves to model chiller efficiency; however, the reviewer chose to add additional empirical models to the study to create a custom baseline energy model. A Gordon-Ng thermodynamic model and a random forest regressor were compared to capture not only system performance but also model uncertainty. We found that the empirical model outperformed the thermodynamic based model by 22% (comparing cross validated R^2 estimates) as shown in Table 2 below.

Table 2. Model performance metrics comparison for site E

Model Performance Parameter	Gordon Ng	Linear Regression	Decision Tree
Model R-squared value (cross validated)	0.72	0.85	0.88
Model R-squared value (test set)	0.70	0.82	0.85
Model RMSE-CV	19.5	14.76	13.12
CV-RMSE predicted vs. actual	0.24	0.05	0.03
R-square predicted vs. actual	0.63	0.83	0.92

Source:

Feature selection was used to rank the independent variables for linear and random forest variables using recursive feature elimination. As expected, the top three variables ranked were chiller plant tons, chilled water return temperature and condenser water supply temperature. The baseline chiller model was used to predict energy consumption in post retrofit scenario to calculate avoided energy savings which were eventually weather normalized to calculate final savings estimates. Without use of feature selection, normalization and random forest, we would not have been able to appropriately capture the change in system performance between baseline and performance periods.

Conventional M&V methodologies can struggle when measures lack established historical data or when savings are too small to stand out in utility data. Thoughtful feature engineering, model design, and model selection can improve accuracy and yield estimates with significantly greater confidence.

Conclusion

As energy systems grow more complex and data-rich, traditional M&V approaches, particularly those based on IPMVP Options A and C, require refinement to maintain accuracy, transparency, and scalability. Conventional methodologies often struggle when measures lack reliable historical data or when savings are too small to be distinguished in utility records. By leveraging onboard monitoring capabilities and applying thoughtful feature engineering, model design, and model selection, practitioners can achieve more accurate and reliable savings estimates, even in complex or data-limited contexts. The case studies illustrate that these approaches are both feasible and necessary to ensure fair attribution of savings, support program cost-effectiveness, and build confidence among stakeholders. As utilities and regulators demand higher precision and transparency, the integration of “M&V 2.0” approaches represents a critical step in aligning evaluation practices with the realities of modern building systems.

References

Mitchell, T. M. 1997. *Machine Learning*. New York: McGraw-Hill.

Samuel, A. L. 1959. “Some Studies in Machine Learning Using the Game of Checkers.” *IBM Journal of Research and Development*, 3(3): 210–229.

Im, P., Joe, J., Bae, Y. J., and New, J. R. 2020. “Empirical Validation of Building Energy Modeling for Multi-Zone Commercial Buildings in Cooling Season.” Knoxville, TN: Oak Ridge National Laboratory.

James, G., Witten, D., Hastie, T., and Tibshirani, R. 2021. “An Introduction to Statistical Learning: With Applications in R.” New York: Springer, 247.

EVO (Efficiency Valuation Organization) 2022. *International Performance Measurement and Verification Protocol: Core Concepts*, Volume I. Washington, DC: EVO.