ENERNOC

White Paper

The Demand Response Baseline

Overview

If only two things matter in Demand Response (DR), they would be:

- How DR resources perform
- How to measure DR performance

It is the second item on this list that we discuss in this paper. Given that the first is largely in the hands of DR participants, DR aggregators, and external factors like the weather, it is the second that should be of utmost concern for policymakers and DR program designers.

Introduction

The measurement and verification (M&V) generally, and the "baseline" more specifically, of DR determines the magnitude of the resource and thus plays an important role in determining the value it has to the electric system. M&V also drives customer compensation for participation, and as a consequence, will influence the number and types of customers for whom the DR program appears attractive. Although there are many methods currently in use, some are much more accurate than others in estimating the fundamental baseline question: what would the customer's load have been in the absence of a DR event?

In 2009, the Federal Energy Regulatory Commission (FERC) signaled interest in developing standards for DR measurement and verification when it noted in its report on the National Assessment of Demand Response Potential, that "development of standardized practices for quantifying demand reductions would greatly improve the ability of system operators to rely on demand response programs" and "central to the issue of measurement is a determination of the customer baseline".[1] FERC tasked the North American Energy Standards Board (NAESB) to develop M&V standards, and in response, NAESB created a glossary of DR-related terms and defined broad types of DR programs and performance methods. This paper continues these efforts by providing definitions, discussions, and recommendations concerning the appropriate application of specific baselines for DR M&V.

EnerNOC used both external and internal resources to inform this paper. First a review of literature on DR M&V was completed to learn from previous baseline studies. Second, we performed quantitative analyses using EnerNOC customer data to examine the merits of different baseline methodologies. This paper is not intended to be a basic primer on demand response, but rather an advanced discussion of baseline methodologies.[2] The majority of baselines discussed herein are applicable to commercial and industrial customers. Demand response programs that primarily include residential customers and their baselines are not the focus of this paper.

> "Good baseline design is driven by adherence to three fundamental pillars: accuracy, simplicity, and integrity."

This paper is organized into three broad sections. Part I discusses the role of baselines in DR programs, outlines the timing of an event, and provides a basic baseline example. Part II includes a description of each of the five baseline methodologies as defined by FERC. Quantitative and qualitative arguments are presented in Part III along with specific recommendations for best practices in the design and application of baselines.

Over the course of this paper, we hope to demonstrate that good baseline design is driven by adherence to three fundamental pillars: **accuracy**, **simplicity**, and **integrity**. While no baseline is perfect, baselines that balance these principles are better than those that do not.

[1] Federal Energy Regulatory Commission. National Assessment of Demand Response Potential, June 2009, p72.

[2] For background on demand response programs, please review Demand Response: A Multi-Purpose Resource for Utilities and Grid Operators, a whitepaper published by EnerNOC and available online.

Section 1. Basics of Demand Response and Baselines

Demand response is a reduction in the consumption of electric energy by customers from their expected consumption in response to an increase in the price of electric energy or to incentive payments designed to induce lower consumption of electric energy.[3] Demand response programs come in a variety of types. Some DR programs are created and run by utilities who work directly with customers to form and execute curtailment plans. Other utility-based DR programs are delivered in the form of dynamic pricing tariffs that encourage customers to reduce their loads during peak times. There are also DR programs that are designed by utilities or system operators in which aggregators recruit customers and take responsibility over customer curtailment actions.

DR programs have different incentive schemes and program objectives. Two of the primary types of incentives are capacity payments and energy payments. Programs provide capacity payments to customers to stand by to be ready to help the grid, either to reduce peak demand or to stabilize the grid during an emergency and prevent blackouts. Energy payments are provided based on the actual energy provided (e.g., not consumed) by a customer over a set period of time during a demand response event. DR programs use these types of payments to incentivize customers to participate. For customers participating in dynamic pricing programs, the incentives are typically represented by rate discounts during the off-peak periods that more than offset the significantly higher rates during the critical peak periods.

Why Baselines Matter

Typical demand response programs rely upon incentivizing energy users based on the extent to which they reduce their energy consumption and therefore require a reliable system to measure energy reduction. For this reason the measurement and verification of demand response is the most critical component of any program. The baseline is the primary tool for measuring curtailment during a DR event.

> *A baseline is an estimate of the electricity that would have been consumed by a customer in the absence of a demand response event.*[4]

Baselines enable grid operators and utilities to measure performance of DR resources. A well-designed baseline benefits all stakeholders by aligning the incentives, actions, and interests of end-user participants, aggregators, utilities, grid operators, and ratepayers. Baselines are a challenging aspect of demand response programs because they must represent what the load would have been if a customer had not implemented curtailment measures. In other word, a baseline is a "counter-factual," a theoretical measure of what the customer did not do, but would have done, had there not been a DR event.

No estimate is perfect, but there are some baselines that are superior to others or best suited to specific programs or customer types. When evaluating a baseline method, it is EnerNOC's experience that three factors are critical above all others – accuracy, simplicity, and integrity.

Accuracy

Customers should receive credit for no more and no less than the curtailment they actually provide, so a baseline method should use available data to create an accurate estimate of what load would have been in the absence of a DR event.

Simplicity

The baseline should be simple enough for all stakeholders to understand, calculate, and implement, including end-use customers. In addition, it should be possible to determine the baseline in advance of or during DR events, so that it can be used to monitor curtailment performance in real time.

Integrity

A baseline method should not include attributes that encourage or allow customers to distort their baseline through irregular consumption nor allow them to game the system.

Balancing these traits is not simple. In some cases, a baseline resistant to manipulation can be so complex as to be unworkable by program stakeholders. On the other hand, the simplest approaches could allow market participants to exploit the baseline in their favor. Therefore, baselines should be evaluated to ensure they provide for all three attributes of accuracy, simplicity, and integrity.

[3] FERC 18 CFR § 35.28 (b)(4)

[4] North American Energy Standards Board. Business Practices for Measurement and Verification of Wholesale Electricity Demand Response. March 16, 2009, p.9. North American Energy Standards Board. Business Practices for Measurement and Verification of Wholesale Electricity Demand Response. March 16, 2009, p.9.

Baseline Basics

Although there are many types of baselines, thanks to the efforts of NAESB, there is now an "official" FERC-approved (and mandated) basic structure that defines how baselines are created and applied in the measurement and verification of demand response. Figure 1 was developed by NAESB to clarify the timeframes within a DR event.

Notice that there are two types of notifications. In some cases, utilities and/or grid operators may provide advance notification to customers or load aggregators when they know that an event will occur or is likely to occur. When it is certain that an event will be called, utilities and/or grid operators notify that the event has been initiated. Aggregators later notify customers of the event at the agreed upon time schedule – this is the deployment of the resources.

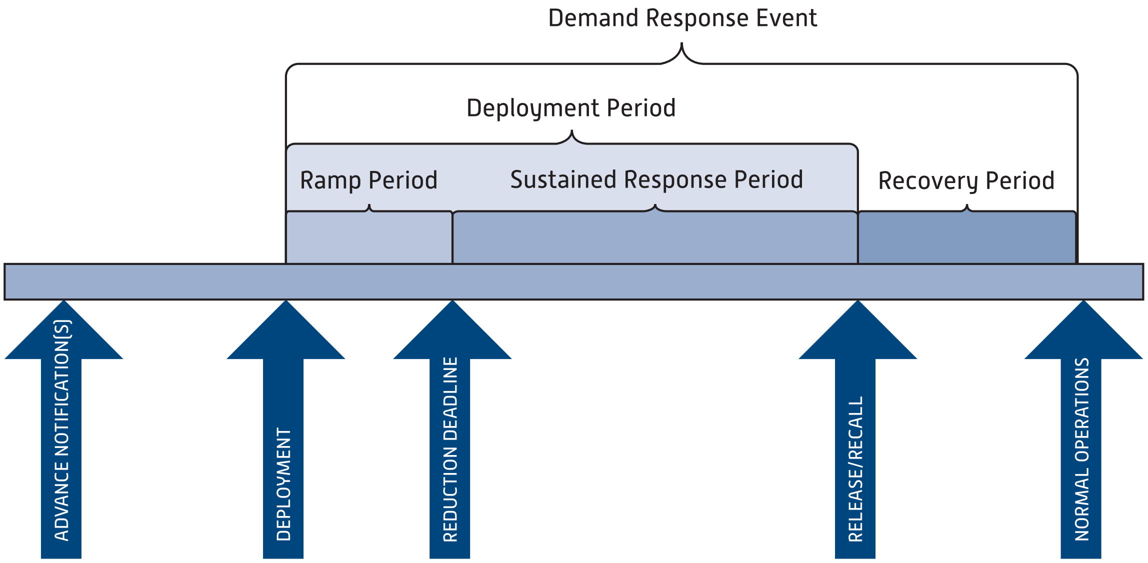
A demand response event has three phases of curtailment.

Phase 1 – The *ramp period*, which begins with deployment, is when sites begin to curtail.

Phase 2 – The *sustained response period*, which is the time period bounded by the reduction deadline and the release/recall, is the time in which the DR resources are expected to have arrived and to stay at their committed level of curtailment.

Phase 3 – The *recovery period*, which occurs after customers have been notified that the event has ended, is the period when customers begin to resume normal operations.

FIGURE 1. TIMING OF DEMAND RESPONSE EVENT[5]

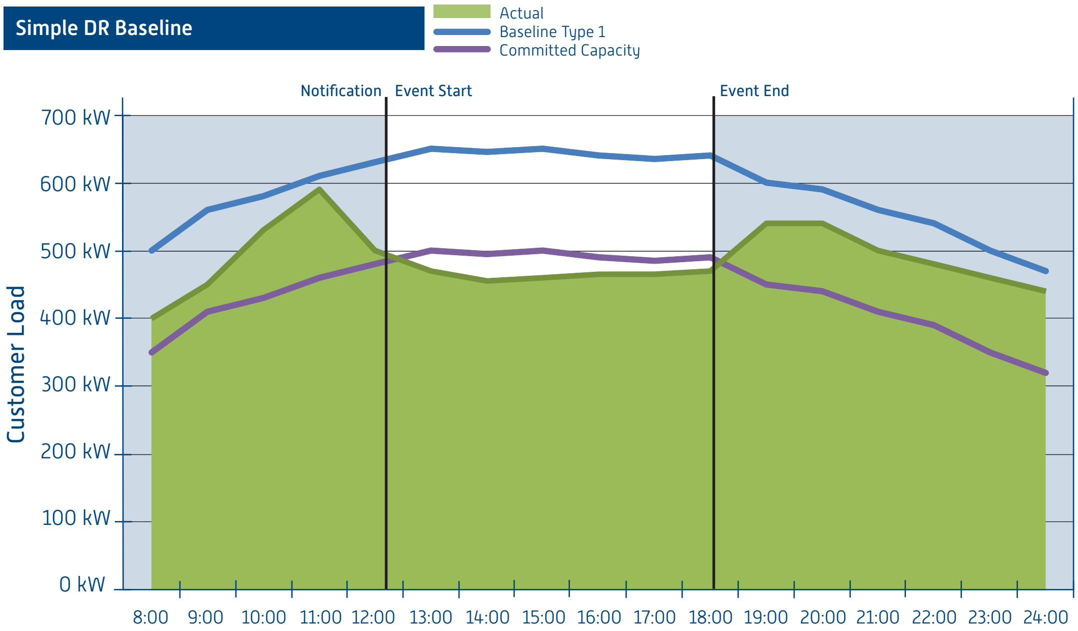


[5] NAESB, p7.

Basic Baseline

Remember that the baseline is the electrical load that the customer would have consumed in the absence of an event. Actual meter data from the period of the event is compared with this baseline to determine the customer's curtailment. Consider the example below in Figure 2. When a customer enrolls in a DR program, engineering specialists working for a utility or aggregator help identify the **committed capacity**—the capacity that a customer will be expected to provide during an event based on the nature of its operations and its curtailment plan. Once a baseline is generated for a customer, a second line can be created to show the committed capacity, or the usage level that a customer must remain at or below during an event. Suppose that a deployment occurs at 11:00 am and the customer begins to decrease energy usage in preparation for the 12:00 pm reduction deadline. Performance can be tracked by comparing the committed capacity (purple line) to the actual meter load (green line). In this case, the customer performed well because during the event the actual meter readings were consistently below the committed capacity level.

FIGURE 2. ILLUSTRATION OF A SIMPLE DR BASELINE



Section 2. Primer on Baseline Types

Programs throughout the United States use a variety of baselines. Some baselines are more appropriate than others based on program type, customer type, and/or program season. There is no perfect baseline—they are all estimates. Factors such as the conditions that trigger a DR event, the frequency of DR events, timing of notification, and duration of event lead to discrepancies between the optimal baseline characteristics for different program types and customers.

In its publication of DR standards, NAESB defined five types of baseline methodologies:

- Baseline Type I - baseline is generated using historical interval meter data and may also use weather and/or historical load data to generate a profile baseline that usually changes hour-by-hour
- Maximum Base Load (also known as Firm Service Level in PJM) – uses system load and individual meter data from the past DR season to generate a flat, constant level of electricity demand for the baseline that the customer must remain at or below
- Meter Before – Meter After – baseline is generated using only actual load data from a time period immediately preceding an event
- Baseline Type II – statistical sampling generates a baseline for a portfolio of customers in the instances where interval meter for all individual sites is not available
- Generation – baseline is set as zero and measured against usage readings from behind-the-meter emergency back-up generators. This type of baseline is only applicable for facilities with on-site generation and is not discussed in this paper.

These baseline methodologies differ in regards to baseline shape, type of data used, timeframe of historical data, and program objective and design. In the following sections, each baseline methodology is explained. Part III addresses the merits and appropriate application of these baselines.

Baseline Type I

> *"A baseline performance evaluation methodology based on a demand resource's historical interval meter data which may also include other variables such as weather and calendar data."*[6]

The Baseline Type I method is the most prominent in demand response programs today. Variations of this method include Averaging, Regression, Rolling Average, and Comparable Day. Characteristics of Baseline Type I methods

- Baseline shape is **the average load profile**
- Utilizes meter data from each **individual site**
- Relies upon **historical meter data** from days immediately preceding DR event
- May use weather and calendar data to inform or adjust the baseline

Averaging Methods

The most widely used Baseline Type I methods are the averaging methods, which create baselines by averaging recent historical load data to build estimates of load for specific time intervals. Averaging methods are often called representative day methods or High X of Y methods.

> *A High X of Y baseline considers the Y most recent days preceding an event and uses the data from the X days with the highest load within those Y days to calculate the baseline.*

High X of Y programs are used throughout the United States. For example, a High 4 of 5 baseline is used in PJM, a High 15 of 20 baseline in Ontario and a High 10 of 10 in California.

> **"Baseline methodologies differ in regards to baseline shape, type of data used, timeframe of historical data, and program objective and design."**

[6] NAESB, p5., p5.

[7] PJM Subcommittee report.

[8] While EnerNOC advocated for a limit to the look-back window to force use of recent days for the baselines, or even event days if there were not enough admissible days within the look-back window, the ISO-NE program rules were changed in a different way to limit participation frequency and thus limit the number of days to be excluded. EnerNOC opposed this blunt approach at the time and has proposed alternative approaches to address the issue of "stale" baselines.

Selection of the number of days to use for a High X of Y baseline is determined by the following considerations:

Look-back Window

The look-back window is the range of days prior to the event day that should be considered in identifying the Y days for a High X of Y baseline. In 2007, the Customer Baseline Subcommittee of PJM conducted a study of baselines, and one parameter they examined was the look-back window. The study concluded that 30 days was too restrictive and that a 60 day look-back window was reasonable and should be used.[7] Today, many programs do not have a restriction on the look-back window; however, it is helpful to have a limit in order to avoid using data that is extremely outdated and thus likely not representative. For example, in 2009, ISO New England (ISO-NE) made a change to their economic demand response program because of evidence that the lack of a look-back window resulted in baselines that used out-of-date interval meter data.[8]

Exclusion Rules

When calculating a High X of Y baseline, certain days prior to the event day are excluded from the Y eligible days, generally because the load on those days is characteristically different from load on regular business days when events occur. It is generally accepted that previous DR event days, holidays, and weekends should be excluded. Many programs have adopted the standard that holidays are those days that the North American Electric Reliability Corporation (NERC) identifies as "Off-peak days."[9]

In addition to the basic exclusions, thresholds and scheduled shutdowns have also been considered. An analysis by PJM in 2008 examined the use of thresholds to aid in selection of the Y days. A threshold of 10% means that a day prior to the event day is excluded from the Y eligible days if the average load during a specified window on that day is less than 10% of the average load of all Y days under consideration over that same time frame. The PJM results showed that a High 5 of 7 day baseline could be improved when a threshold of 25% was used rather than 10%. A study by Lawrence Berkeley National Labs (LBNL) recommended that scheduling information related to shutdowns and large swings in energy be included to help inform baseline predictions. In particular, LBNL recommended that there be a way to capture when facilities are closed on Mondays or during the summer and to use that information when calculating baselines.[10] The exclusion of DR event days, holidays and weekends is necessary, but further exclusions such as load thresholds or scheduled shutdowns can greatly increase the complexity of baseline calculations.

Relationship Between X and Y

Once a group of prior days is identified as the Y days, that group of days is narrowed down to a subset of X days in order to obtain a better representative group of days. This subset of days should be formed based on the nature of the program. For example, a demand response event within a summer emergency DR program is often called on a day when load is expected to be high, usually driven by extreme weather conditions. Not all of the eligible Y days, however, will have been days with high load. Thus, an unadjusted baseline that uses data from all Y days will include a number of non-event-like days and the baseline will consistently understate the participant's true baseline, reducing the incentive to participate while challenging the accuracy of the program. To avoid this baseline understatement, many programs remove the days with the lowest load levels. Alternatively, operators of peak shaving programs that operate year round may have observed that events do not always overlap with high load days. Thus, those programs may choose to use a Middle X of Y baseline rather than a High X of Y baseline, in order to capture a more appropriate middle level of load during the fall and spring seasons. Another approach that has merit for other reasons is the use of an adjustment (see Baseline Adjustments section below on page 8) to the High X of Y, or Middle X of Y baselines.

Time Intervals

Most programs capture frequent intervals of data. This captures greater detail around the load behavior of customers. In most analyses of baselines, hourly load data was used, simply because processing 5-minute interval data for hundreds of customers and lots of baselines was an unnecessary logistical strain.

[9] See document: www.nerc.com/docs/oc/rs/Additional_Off-peak_Days.doc

[10] Lawrence Berkeley National Laboratory (LBNL). Estimating Demand Response Load Impacts: Evaluation of Baseline Load Models for Non-Residential Buildings in California. January 2008, p27.

SPOTLIGHT: Technical Computations of High X of Y Baseline

Consider fictional customer, TechBiz, a commercial customer that operates an office building and minor manufacturing processes. Like most commercial customers, TechBiz increases energy usage in the morning when employees arrive at work and decreases usage in the afternoon when employees leave. Occupancy is typically very low during the weekends, so the corresponding energy usage is also quite low during these periods. Assume that an event has been called for Wednesday, July 22nd and that the program in which TechBiz participates uses a High 5 of 10 baseline.

Figure 3 shows load data from the meter at TechBiz. Each bar represents the average load value (in kW) from 12 pm –8 pm for TechBiz. Before searching for the Y most recent days, weekends, holidays and prior event days (colored blue, purple, and green, respectively) must be excluded. The remaining ten most recent days, the yellow bars, are then ranked by the sum of the load usage over the hours of the sustained response period. The five days with the highest load (July 9, 10, 15, 18 and 19, shown with shadowed yellow bars) are selected to calculate the baseline. The baseline for each time interval, in this case each hour, is determined by averaging the load on those five days for each hour.

Baseline Adjustments

As mentioned, the subset of X days is designed to consist of days similar to the event day. The conditions on the event day, however, are often different from prior day conditions, especially for customers with weather-sensitive loads that increase during extremely hot and/or extremely cold conditions. Programs that are triggered by peak demand conditions or emergencies caused by generation outages often coincide with days of extreme weather temperatures. For this reason, High X of Y baselines are often adjusted. A baseline adjustment is sometimes called a "day of adjustment" because the adjustment is made based on data from the day of the event. Even for programs that are not likely to be called on days of abnormally high loads, adjustments help in situations where load is lower or higher than it has historically been, and the baseline doesn't accurately capture the load behavior immediately prior to the event on the event day.

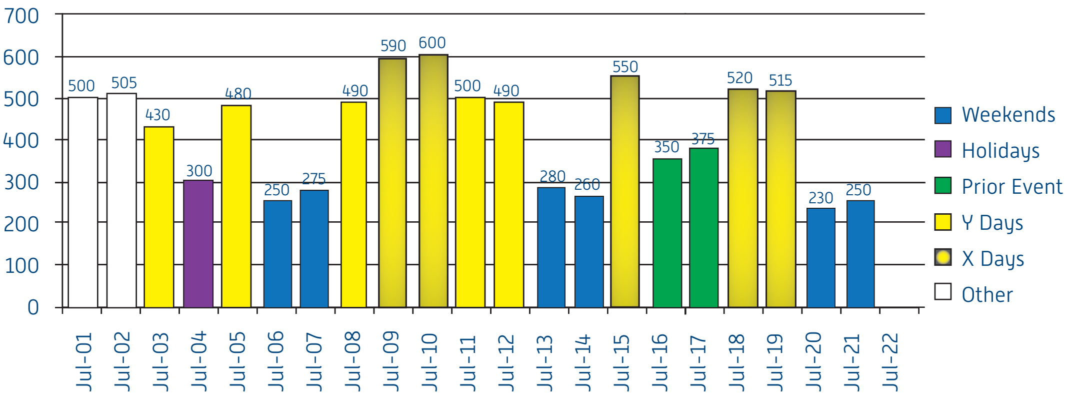
> *"An **adjustment** to a High X of Y baseline is necessary to more accurately reflect load conditions of the event day."*[11]

An adjustment is defined by the time frame that is used to make the adjustment and by the choices to use adjustments that are scalar or additive, capped or uncapped, and symmetric or asymmetric.

Timing & Duration

Most baseline adjustments use a timeframe of 2-4 hours prior to the event. More than 1 hour is needed to be representative of the difference and 4 or more hours may consider conditions too far away from the event to be representative. Actual load over this time period is compared to the load estimated by the baseline over the same time period and is used to calculate the appropriate adjustment. It is optimal for the adjustment

FIGURE 3. ILLUSTRATION OF A SIMPLE DR BASELINE



[11] NAESB, p10.

> "In order to limit the magnitude of any adjustment, some programs use a cap… In EnerNOC's experience, capped adjustments can penalize customers on days of extraordinarily high load."

to use load values from time intervals preceding deployment. For example, if load is used during the ramp period, then the adjustment could penalize a customer for early curtailment or allow a customer to game the system by increasing load temporarily. Both these issues compromise the integrity and accuracy of adjustments, but can be avoided by using a timeframe for the adjustment that precedes notice of the event start. For example, if an event starts at 12:00pm and customers will be notified at 11:00am, then the load from 8am–11am could be used to calculate an adjustment.

Scalar vs. Additive

Adjustments can be calculated using a scalar or an additive factor. The scalar technique is based on a percentage comparison. If load on an event day prior to notification is 30% above the calculated baseline, then each time interval of the baseline would be 130% of the calculated baseline. The additive approach instead calculates the actual demand difference in kW. If load during the calculation period is 50 kW above the calculated baseline, then 50 kW is added to each interval in the actual event baseline. EnerNOC did not conduct an analysis of whether scalar or additive adjustments lead to more accurate results. In EnerNOC's experience, scalar adjustments can produce highly erratic and exaggerated results when load during the adjustment window is very low. Accordingly we recommend the use of an additive adjustment.

Capped vs. Uncapped

In order to limit the magnitude of any adjustment, some programs use a cap. For example, a customer with 100 kW baseline exhibits demand of 130 kW prior to event notification. Using an additive adjustment, the customer baseline throughout that day's event would be increased by 30k W. If the program uses a 20% cap, however, then the additive adjustment would be limited to 6 kW. In EnerNOC's experience, capped adjustments can penalize customers on days of extraordinarily high load, as discussed on page 15.

Symmetric vs. Asymmetric

It is important to consider whether adjustments reflect demand conditions symmetrically (baseline adjusted up and down) or asymmetrically (baseline only adjusted up). The symmetric approach considers that day-of conditions can have a real impact on customer demand in both directions and therefore symmetric adjustments can maximize the accuracy of a baseline calculation. However, a symmetric adjustment can permit downward adjustments that could have damaging unintended consequences.

For example, a customer may decide to shut down a product line after a batch is complete, because the customer knows the reduction deadline is approaching. If the baseline uses meter data from after the production line has been shut off in order to compute the adjustment, then the baseline of that customer could drop and misrepresent the expected load conditions in the absence of an event. For this reason, program designers must take careful consideration to avoid any overlap of the timeframe used to calculate the adjustment and the ramp period.

Regression

Another variant of the Baseline Type I is a regression baseline. This baseline is built using a customer-specific regression analysis to estimate load based on prior load behavior, weather conditions, calendar data, system demand, and time of day. Regression analysis may be the most accurate of baseline methodologies because it takes into consideration more variables that influence load. Over the last ten years, numerous groups have compared the merits of regression baselines to High X of Y methods. A study by LBNL found that High X of Y methods work better than regression methods for high load variability customers.[12] A previous study for California Energy Commission (CEC) in 2003 discovered that High X of Y methods perform close to weather regression models.[13] Later in 2009, however, an analysis by the Association of Edison Illuminating Companies (AEIC) showed that regression methods outperform High X of Y methods.[14] Quantum Consulting recognized that regression may be a more accurate method, but not practical for real-time use during events.[15] Typical explanatory variables used in regression models tested in these studies include Average kW, Cooling and/or Heating Degree Days, Day Type Indicators, and Day of Average kW.

Regression baselines are complex to calculate and as mentioned they require load, weather, and day-type data. They may rely on interval meter data from an entire summer to estimate load during event days of that summer. In this case, it is not possible to calculate a baseline in real time during an event, since the regression equation can only be created at the end of the summer. For customers and aggregators, it is important to present the baselines during an event, because it shows a customer whether it is meeting curtailment expectations. Furthermore, regression analysis can also delay post-event performance evaluation and hurt customer satisfaction when results cannot be delivered in a timely fashion. Regression baselines sacrifice too much simplicity for accuracy; therefore, they are not a preferred M&V method for any DR programs.

Other Baseline Type I Methods

Two other Baseline Type I methods are Comparable Day and Rolling Average. The **Comparable Day** method allows an aggregator to find a day that is similar to the event day and use the load of that similar day as the baseline for the actual event day. This method still uses historical meter data, but unlike the Averaging methods, it uses only data from one day, rather than from multiple days. Two challenges with Comparable Day are 1) it is not possible to know the baseline during the event which could impede meeting curtailment goals, and 2) there are no objective criteria for selection of the day which makes it difficult to assess the appropriateness of a comparable day.

The **Rolling Average** baseline uses historical meter data from many days, but gives greater weight to the most recent days. The baseline relies on a greater number of data points, which could improve accuracy for a customer who has similar load patterns and levels throughout the year. For customers whose energy usage fluctuates between seasons, however, the rolling average may not be the best method. For example, suppose a customer is a ski area with ski lifts that are closed down for most of the summer. In the winter, the ski lift operates for 10 hours each day. If an event was called at the beginning of the ski season, a Rolling Average baseline would reflect the summertime usage and might be too low for the customer to receive any credit for curtailment.

Maximum Base Load

> *"A Maximum Base Load is a performance evaluation methodology based solely upon a Demand Resource's ability to reduce to a specified level of electricity demand."*[16]

Maximum Base Load (MBL) methods identify the maximum energy usage expected of each customer and then set a specific level of electricity usage that is equal to the maximum level minus the committed capacity of the customer. MBL methods are sometimes referred to as "drop to" methods because a customer must drop to a specific level of usage during an event. In comparison, most Baseline Type I methods are referred to as "drop by" because the customer knows the amount of committed capacity that they must drop, but the level of usage is not necessarily a constant level. The MBL is an example of a static baseline, because it remains at one level, as compared to a Baseline Type I method that generates a dynamic, changing profile of the load throughout the hours of the day. Note that with an MBL baseline, it is entirely possible for a customer to "perform" by doing nothing all, so long as its load is already at or below the "drop to" level.

Characteristics of Maximum Base Load methods

- Baseline shape is **static**
- Utilizes meter data from each **individual site** and from the **system**
- Relies upon historical meter data from **previous year**

Two well known examples of Maximum Base Load methods are the Average Coincident Load (ACL) used in the New York Independent System Operator (NYISO) Special Case Resources (SCR) program and the Peak Load Contribution (PLC), which is part of the Firm Service Level (FSL) option in PJM's Emergency Load Response Program (ELRP) program. Both methods identify the peak hours of the previous year (across a subset of hours) and then use the load on those hours to create an average maximum load for each customer. This maximum load becomes the baseline for all hours of the current summer. The main difference between the two methods is the manner in which peak hours are identified.

[12] LBNL, p26.

[13] California Energy Commission (CEC). Protocol Development for Demand Response Calculation – Findings and Recommendations. Study conducted by KEMA – XENERGY. February 2003, pX-1.

[14] AEIC Load Research Committee (AEIC). Estimation Errors in Demand Response with Large Customers. November 2009, p8-15.

[15] Southern California Edison Company (SCE). Evaluation of 2005 Statewide Large Nonresidential Day-Ahead and Reliability Demand Response Programs. Study conducted by Quantum Consulting Inc. April 2006, p7-106.

[16] NAESB, p5.

Coincident vs. Non-coincident

An MBL baseline can be either coincident or non-coincident. A coincident baseline uses peak hours of the summer that are chosen based on system load peaks. A non-coincident baseline also uses peak hours, but they are determined by individual load behavior and not by the system load. This means that the hours that contribute to the non-coincident baseline vary among customers.

Meter Before/Meter After

> *"A performance evaluation methodology where electricity consumption or demand over a prescribed period of time prior to Deployment is compared to similar readings during the Sustained Response Period."*[17]

In an ancillary services event, the minimal notice and reduced event durations create a set of circumstances that require a unique baseline calculation. Generally, an ancillary services event is intended to reduce load on the grid at that moment, for a short period of time, rather than to reduce a dynamic load profile likely to fluctuate over time.

Characteristics of Meter Before/Meter After methods:

- Baseline shape is **static**
- Utilizes meter data from each **individual site**
- Relies on **small day-of time interval** of historical meter data

The demand response baselines and programs discussed thus far have been focused on emergency and energy services where customers participate for 4–8 hours at a time. In ancillary services, however, the duration is much shorter, usually 10 minutes to two hours. For this reason ancillary programs typically use meter before-meter after baselines.

Baseline Type II

> *"A performance evaluation methodology that uses statistical sampling to estimate the electricity consumption of an Aggregated Demand Resource where interval metering is not available on the entire population."*[18]

Most baselines are created using historical meter data from the individual site of the customer. There are instances, however, where data from individual sites is not available, but instead data from a meter that aggregates or is representative of several sites is available. In these cases, the meter data can be used to create a baseline for a group of sites and then a method used to allocate load to specific sites. For example, consider a group of sites that are homogenous with similar load behavior. A Baseline Type II method could meter a few of the sites in order to develop an average load estimate per site and then use that to allocate load from the aggregated baseline.

In DR programs with commercial and industrial customers, which are the focus of this white paper, Baseline Type II methods are not common, because most sites either have or can be cost-effectively equipped with interval meters. The Baseline Type II method is more often used in residential DR programs, where it has been cost-prohibitive to install interval meters at every house. As deployment of residential interval meters increases, however, the need for Baseline Type II methods will likely decrease.

SPOTLIGHT: Technical Computations of a Maximum Base Load

Consider again, TechBiz,and assume that TechBiz is now in a program that uses a MBL method. The PLC is the average load of the five peak hours during the summer months June through September of the prior year, coincident with the hours on which the system (the PJM RTO) had the highest load. Only one hour per day can be used as a coincident peak hour for the PLC and weekends and holidays are excluded when determining peak load hours. As an example, Figure 4 shows PJM's five highest summer 2008 peak hours. Based on the historical corresponding load information for TechBiz, its 2009 PLC would be 600 kW, which is the average of the five hourly coincident values.

FIGURE 4. HISTORICAL LOAD DATA FOR THE PLC

System Peak Rank	Hour	Date	TechBiz Load
1	4-5pm	June 09, 2008	590 kW
2	4-5pm	July 17, 2008	640 kW
3	4-5pm	July 18, 2008	615 kW
4	4-5pm	July 21, 2008	580 kW
5	4-5pm	June 10, 2008	575 kW

[17] NAESB, p5.
[18] NAESB, p5.

Section 3. Comparison of Baseline Methods

If no baseline is perfect, then what baseline or baselines are best? This question has intrigued and plagued demand response program designers and participants for years. Past studies have examined the quantitative and qualitative characteristics of the methods. EnerNOC believes that the three pillars of accuracy, simplicity, and integrity outlined above can serve as the foundation for assessing the appropriateness of baseline methods.

To assess the merits of different baseline methodologies, EnerNOC conducted several studies in 2010. The results of these analyses are summarized below as well as recommendations for the appropriate creation and application of baselines they suggest.

Maximum Base Load Analysis

In 2010 EnerNOC conducted a study to compare the merits of non-coincident and coincident MBL methods. The study used customer data and baselines from the PJM ELRP, because this program includes customers that use a High X of Y baseline as well as customers that use a Maximum Base Load baseline. In PJM the MBL baseline is referred to as the Peak Load Contribution (PLC).

Data was analyzed from 120 customer sites in PJM and across five summer 2009 event-like days, identified based on weather conditions and system load. Event-like days are days when events did not occur, so that normal meter readings from each site are available and can be compared to the calculated baseline to assess the accuracy of the estimate. Four baselines were examined: Coincident MBL, Non-coincident MBL, High 4 of 5, and High 4 of 5 Adjusted (additive only).[19] These baselines were calculated using methods identical or similar to those used in the PJM ELRP. Results were assessed using the metric of Median Percent Error = (Baseline$_{h1}$ – Meter$_{h1}$)/Meter$_{h1}$, which compared the baseline to the actual meter data from the event-like day. Figure 5 displays the results.

Conclusions

Baseline Type I Methods exhibit higher accuracy and lower variability compared to MBL methods. The Baseline Type I methods (High 4 of 5 and High 4 of 5 Adjusted) both have lower median percent errors than the MBL baselines. Indeed, the Baseline Type I methods have errors that were close to or at zero. Furthermore, the variability in median percent errors of the Baseline Type I was significantly smaller than the variability in the MBL baselines as shown by the smaller width of the bars.

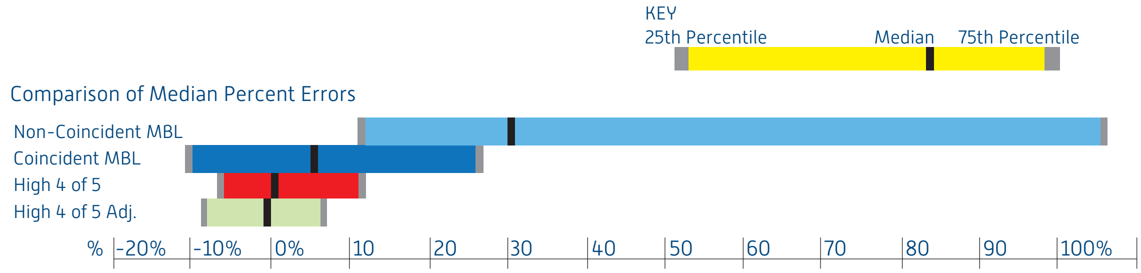
Both MBL baselines overstate meter load and have high variation in errors. Results show a 5% over-bias with the Coincident MBL and a 30% over-bias with the Non-Coincident MBL. While the 5% over-bias is a low overstatement, both MBLs also have a high variability in the median percent errors, as shown in Figure 5.

Non-coincident baseline increases overstatement of expected meter load by six times that of Coincident baselines. Non-coincident peak days set the baseline artificially high. This allows the baseline to almost always be higher than load and thus a baseline based on non-coincident peak days will overstate performance.

Implications & Recommendations

For most customers, it is better to use a Baseline Type I method rather than an MBL method. Although an MBL baseline offers simplicity, this benefit is outweighed by

FIGURE 5. GRAPHICAL RESULTS OF MBL ANALYSIS



[19] Further details can be found in Analysis of Baseline Methodologies and "Best Practice" Recommendations, Brianna Tufts and Aaron Breidenbaugh, presented at the Association of Energy Service Professionals National Energy Service Conference (Tucson, AZ), February 2009.

its poor accuracy. Aggregators can produce systems that reduce the complexity of Baseline Type I methods to a manageable level and thus EnerNOC believes that Baseline Type I methods are preferable for most customers.

In a limited number of situations, however, MBL methods may be more appropriate than Baseline Type I methods. MBL methods are straightforward and simple to implement, given that they only need to be computed once a season, and they provide little opportunity for participants to distort their baselines. Two instances where MBLs are preferred are as follows:

1. **Customers with volatile loads.** Some customers have very volatile loads throughout the day, but enroll in demand response programs because they can easily curtail large amounts of energy usage. A Baseline Type I method will not be able to accurately forecast load for a customer with volatile load patterns, and so it is better to use an MBL method to set the baseline and expectations for curtailment.

2. **DR programs that are intended to ensure that load does not exceed levels used for planning and for which real-time load reductions are irrelevant.** Demand response is one tool that may be used to help grid operators and utilities manage load throughout the network, in particular by ensuring that overall system load, or load behind a specific substation, does not exceed a predefined threshold. In this instance, action during an event needs to guarantee that usage does not exceed a specific level. Baseline Type I methods can guarantee that a certain amount of capacity is available in real time, but MBL methods are more appropriate for ensuring the load stays at or below a set level.

In those situations **where an MBL method is more appropriate, a coincident baseline should be used**. The EnerNOC study described above, as well as one conducted by NYISO in 2010, both confirm that non-coincident baselines drastically overstate performance in comparison to coincident baselines and Baseline Type I methods.[20] Evidence shows that compared with a non-coincident method, the coincident baseline better predicts the collective load of enrolled resources during peak periods, as well as provides a better estimate of real-time load reduction.

High X of Y Analysis

In summer 2010, EnerNOC conducted a quantitative study to examine the relationship between X and Y in High X of Y baselines.[21] Many of the previous studies on baselines, except for the PJM analysis, concentrated on only a handful of baselines and used data from a small sample of customers. While it is understood that X is a subset of Y, there is less agreement about the appropriate size of X relative to Y. For example, current programs use High 3 of 10, High 5 of 10, High 4 of 5, and High 15 of 20 baselines. This analysis looked at unadjusted and adjusted High X of 5 and High X of 10 baselines because they are the most prevalent in demand response programs. The objective of studying all iterations of the baselines was to search for trends in High X of Y baselines and to glean best practice recommendations not by finding a single, perfect baseline, but by understanding how baseline methods compared given the relationship between X and Y and the adjustment cap.

The basic design of this study included:

- 306 sites from NYISO, ISO-NE, PJM, SCE, PG&E, and ERCOT
- 540 baselines per site
    - High X of 5 and High X of 10 for every value of X
    - 1 unadjusted and 5 adjusted baselines, including
        - Adjustment 1 - 3 hours duration, starting 4 hours before the event, with no cap and 20% cap
        - Adjustment 2 - 2 hours duration, starting 3 hours before the event, with no cap, 20% cap, and 40% cap
    - All adjustments are symmetric additive. Adjustment 1 is 3 hours duration, starting 4 hours before the event, while Adjustment #2 is 2 hours duration, starting 3 hours before the event.
    - Baselines generated for 3 event-like days in 2008 and 2009[22]
    - Evaluated summer season - June through September
- Median Percent Error calculated for each baseline across all customers where Percent Error = $(\text{Baseline}_{h1} – \text{Meter}_{h1)}/\text{Meter}_{h1}]$

Figures 6 and 7 show the results for High X of 5 for 2008 and 2009, respectively, while Figures 8 and 9 show High X of 10 results, also for 2008 and 2009.

[20] New York Independent System Operator (NYISO). Proposal for SCR Baseline Alternative. Presented at Installed Capacity Working Group. October 29, 2010.

[21] Further details can be found in Analysis of Baseline Methodologies and "Best Practice" Recommendations, Brianna Tufts and Aaron Breidenbaugh presented at the AESP National Energy Service Conference (Tucson, AZ), February 2009.

[22] Event-like days selected based on highest system load occurrences.

FIGURE 6. CHANGE IN MEDIAN PERCENT AS X VARIES SUMMER 2008 – HIGH X OF 5

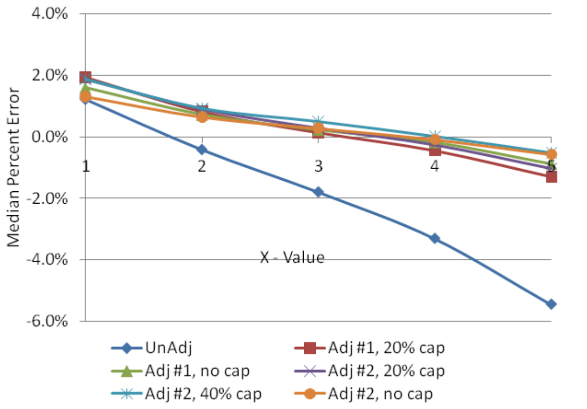


FIGURE 7. CHANGE IN MEDIAN PERCENT AS X VARIES SUMMER 2009 – HIGH X OF 5

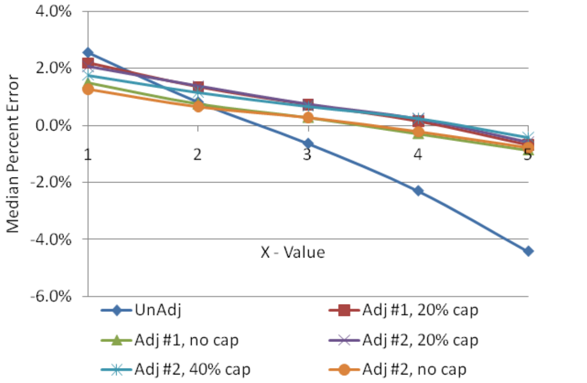


FIGURE 8. CHANGE IN MEDIAN PERCENT AS X VARIES SUMMER 2008 – HIGH X OF 10

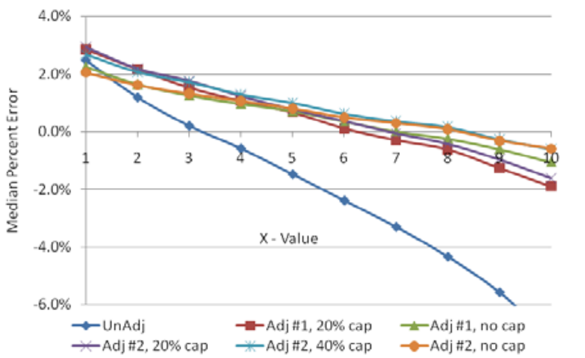
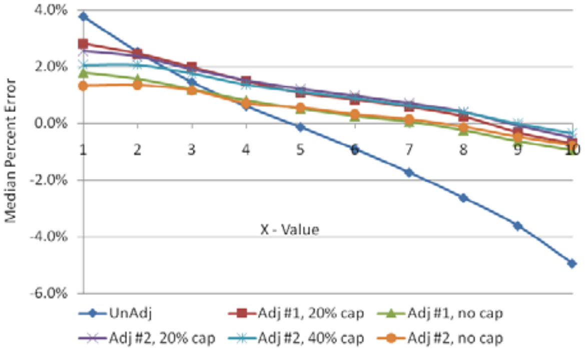


FIGURE 9. CHANGE IN MEDIAN PERCENT AS X VARIES SUMMER 2009 – HIGH X OF 10



Conclusions

Unadjusted baselines increasingly understate the load as X approaches Y. As the X value increases, the median percent error drops steadily, reaching a low of -7.2% for High 10 of 10 in 2008.

Choice of X relative to Y is less critical when using an adjustment. All of the adjusted baselines have median percent errors roughly between +3% and -2%. The slopes of the lines for the adjusted baselines are less steep than the slopes of the unadjusted baselines, indicating less variation.

Adjustments with no cap appear to be the least biased methods. In all four scenarios, the adjustments with no cap have the median percent errors that on average are closest to zero for all values of X, indicating these methods least over- or under-estimate load.

There appears to be a range of values X/Y that minimizes bias. Based on the limited population and years studied, it appears that, for all ranges of adjustments studied and all values of Y, error is minimized by (0.4) < (X/Y) < (0.8).

Implications & Recommendations

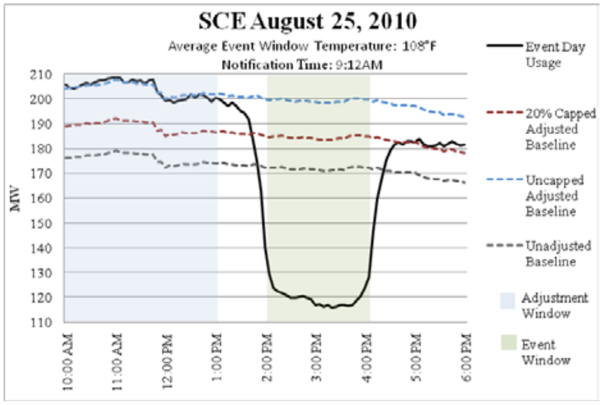
High X of Y Baselines should always be adjusted. While historical meter data can create a reasonable baseline, it is clear that incorporating load information from directly before the event can improve the accuracy of the baseline. Additionally, over recent years, many studies have examined the merits of baseline adjustments. Studies by LBNL, CEC, and AEIC all found that use of an adjustment improves accuracy and reduces bias.[23] LBNL and CEC both recognized that an adjustment could penalize a customer if the adjustment window overlapped with early curtailment actions.[24] In this case, the meter readings would be below normal and the adjustment would shift the baseline downward too much. This would result in a smaller curtailment measurement that underestimated actual performance.

[23] LBNL, p25. CEC, p6-12. AEIC, p20.

[24] LBNL, p18-19. CEC, p6-10.

Adjustments should not be capped. The analysis above shows small differences between capped and uncapped baselines among all customers. Caps can have severe consequences for individual customers. Figure 10 shows an event in California in summer 2010. The program in California used a symmetric baseline adjustment with a 20% cap. In this example, the customer had abnormally high load the morning of an event, but because the adjustment was capped, the baseline could not be adjusted properly to reflect day-of load conditions. Due to this technicality, the customer had trouble reducing to the expected level, despite managing to drop 4 MW. Given that the analysis does show slightly higher accuracy for uncapped adjustments and to avoid glitches such as in Figure 10, adjustments should not be capped.

FIGURE 10. EXAMPLE OF CALIFORNIA DR CUSTOMER WITH UNUSUAL DAY-OF LOAD CONDITIONS



SPOTLIGHT

Interestingly, the results of the EnerNOC analysis are similar to findings of the recently released study, *PJM Empirical Analysis of Demand Response Baseline Methods*, a 300-page analysis that is perhaps the most comprehensive examination of DR customer baseline (Baseline Type-I) alternatives. [25]

The PJM study implicitly endorsed EnerNOC's three pillars of accuracy, simplicity, and integrity. Its specific objectives were to assess the accuracy and bias of a variety of Baseline Type-I methods and to determine the feasibility (or simplicity) of implementing each. With regard to integrity, the PJM study states that "any CBL can be manipulated to the market participant's economic advantage" and recommends" that rules be established to identify and mitigate this behavior."

EnerNOC recommends against regression approaches and found that High X of Y approaches provide a good balance between accuracy and simplicity. The PJM study also concluded that the High X of Y and regression methods it evaluated offered similar accuracy across all segments, and thus regression approaches were not recommended given their greater complexity and thus higher administrative costs than High X of Y methods.

EnerNOC recommends that High X of Y baselines should always be adjusted. While historical meter data can create a reasonable baseline, EnerNOC finds that incorporating load information from directly before the event will improve the accuracy of the baseline.

In particular, unadjusted baselines increasingly understate the load as X approaches Y. PJM's study also concluded that a same-day adjustment "has superior performance to an unadjusted" baseline.

Adjustments should not be capped. Some methods limit the percentage by which a High X of Y baseline can be adjusted. EnerNOC's analysis shows significant differences between capped and uncapped baselines among all customers. In some cases, when customers are experiencing abnormally high usage on an event day, caps may so limit the ability of adjustments to reflect the customer's usage that even if a customer significantly curtails, it may not meet its commitment based on the under-adjusted baseline. The PJM study did not support the use of a capped adjustment. In fact, the study did not even bother to address the merits of using a cap, finding such capping to be "idiosyncratic."

EnerNOC's analysis showed that **Maximum Base Load Methods can overstate baselines by a significant amount and have high variation in errors**. Although MBL methods score points for their simplicity, EnerNOC believes that High X of Y averaging methods are preferable in most cases due to their greater accuracy in predicting real-time future load. The PJM study stated that MBL methods should be the exclusive approach to measuring DR performance in capacity programs, because that is PJM's policy position. PJM subscribes to the view, noted on page 13, that its capacity "DR programs…are intended to ensure that load does not exceed levels used for planning and for which real-time load reductions are irrelevant". EnerNOC is disputing this formulation at FERC. In the PJM study no analysis was conducted comparing the relative accuracy of MBL and Baseline Type-I approaches.

[25] PJM Empirical Analysis of Demand Response Baseline Methods, KEMA, Prepared for the PJM Markets Implementation Committee, April 20, 2011.

The ratio of X:Y should be more than 0.4 and less than 0.8. The X of Y Analysis used data from two different years and from over 300 customers, in order to gain sufficient amounts of data points to outweigh outliers and confounding data. Accuracy is a central pillar of baseline methodology, and there is clear evidence that High X of Y baselines have the lowest median percent errors when (0.4) < (X/Y) < (0.8). Thus, High 3 of 10 baselines are less accurate compared to High 5 of 10 baselines, and High 7 of 10 baselines are preferred to High 10 of 10 baselines.

Asymmetric and Symmetric Baseline Adjustments

The High X of Y Analysis included baselines with symmetric additive adjustments, because symmetric baselines are more widely used than asymmetric baselines. Symmetric adjustments can shift a customer's load both up and down, and as long as they are uncapped, they should be able to best utilize day of information to enhance the accuracy of the baseline. Thus they are likely more accurate than asymmetric adjustments. There are some practical reasons, however, that make asymmetric adjustments simpler and more customer friendly.

Implications & Recommendations

Symmetric adjustments are most appropriate for DR programs not geared for extreme load conditions. Symmetric adjustments are most appropriate for programs in which events are less likely to occur on days of extreme load conditions. For example, in programs that occur during these times, the event day may not be expected to have significantly different loads from day to day. Thus there is an equal chance that an unadjusted baseline could be lower or higher than actual load prior to an event, in which case a symmetric adjustment would be appropriate.

Asymmetric adjustments are better suited for summer and winter programs. These adjustments properly align incentives of participants with objectives of demand response programs, and consequently, these often rate higher from an integrity perspective. A simple example of this would be an extreme weather event where the local government is issuing pleas for load reductions. With a symmetric adjusted baseline, DR participants are in danger of reducing their baselines by responding to the request, and as a consequence, could be subject to lost revenues or even program penalties. Another example would be when DR events have been called on two consecutive summer afternoons, and a DR event on the third day (the hottest of the week) is all but inevitable. In this scenario, the customer might need to start up operations during the baseline adjustment period of the third day just to avoid baseline compromise when, given an asymmetrically adjusted baseline, they could have just cancelled the whole shift and not worried about baseline erosion.

> "Asymmetrically adjusted baselines can help avoid situations where energy users feel punished for taking logical steps to reduce consumption at times of grid stress."

There are ways to reduce any negative repercussions from early curtailment actions, but the best way to prevent them is to use asymmetric adjustments, which will not adjust the baseline down as a consequence of the actions mentioned earlier. Asymmetrically adjusted baselines can help avoid situations where energy users feel punished for taking logical steps to reduce consumption at times of grid stress.

Adjustment windows should not overlap with the ramp period. A common concern in demand response programs is how to avoid penalizing customers for early curtailment. Program operators care most about participants meeting curtailment expectations during the sustained response period and they understand that participants begin to curtail during the ramp period in order to adequately curtail their consumption levels by the reduction deadline. In order to avoid a conflict, the adjustment window (the time interval when meter data is compared to the baseline and that difference is used to adjust the baseline), should come before the ramp period and the two periods should not overlap. In order to facilitate this, the start time of an adjustment window should be based on the deployment time, providing it is sufficiently close to the reduction deadline. Most programs choose to use an adjustment window of 2–3 hours in duration. Thus an acceptable adjustment window could occur 3 hours before deployment with 2-hour duration.

Treatment of Baseline Aggregations

Most programs use interval metering from the customer site to generate a baseline for that customer. Some programs, however, take a different approach and use portfolio baselines.

> *Portfolio baselines are created by first using the aggregate meter data of all customers to generate a baseline for the aggregation and then using individual meter data to assign baselines to each customer.*

Note that portfolio baselines are not an example of Baseline Type II, because Baseline Type II is used where individual metering is not available, but portfolio baselines do have access to individual metering. Consider the example in Figure 11 of a small program with three participants.[26] The program uses a High 5 of 10 Baseline and interval meter data from the last ten days prior to the event day is displayed. Out of the past 10 days, the highest five days of the portfolio are Days 1, 2, 4, 5, and 8 (shown in dark blue). Each participant, however, has a different set of days that represent its highest five days (shown in light blue).

FIGURE 11. COMPARISON OF INDIVIDUAL AND PORTFOLIO BASELINES (LOAD IN KW)

	Participant 1	Participant 2	Participant 3	Aggregate Load
Day 1	200	65	100	365
Day 2	200	65	80	345
Day 3	200	65	70	335
Day 4	200	130	80	410
Day 5	200	130	60	390
Day 6	0	130	80	210
Day 7	0	130	110	240
Day 8	150	130	100	380
Day 9	150	65	125	340
Day 10	150	65	100	315
High 5 of 10, Portfolio	190	104	84	378
High 5 of 10, Individual	200	130	107	437

If a portfolio baseline were used, the baseline for the aggregate load would be 378 kW, but if individual baselines were used, then the aggregate load baseline would be 437 kW. In this example, using individual baselines generates baselines that are higher than baselines generated using the portfolio method. This will always hold true when the High X days of the system are not the same High X days of every participant.

Implications & Recommendations

How do program designers choose between portfolio and individual baselines? They must consider the objectives and characteristics of the program and the customers.

Individual baselines should be used to preserve accuracy.

As seen in the diagram above, the portfolio baselines are much lower than the individual baselines, and under-represent estimated load conditions.

Individual baselines should be used to preserve integrity.

Portfolio methods provide decreased certainty to participants and skew performance calculations. Each site may not necessarily be rewarded correctly for their curtailment, because the aggregator cannot easily reconcile information of individual sites' curtailment actions with curtailment results of the portfolio.

Some demand response programs use portfolio baselines, even though individual baselines are comparably simple to administer. Most notably, a portfolio High 3 of 10 Baseline was used in California in recent years. It was recently changed to a High 10 of 10. While the program still defines it as a portfolio baseline, it is essentially both an individual and portfolio baseline because the High 10 of 10 days are the same for all participants and the system. Given that portfolio baselines are not necessary and that they almost always understate curtailment efforts, it is generally appropriate to use individual baselines rather than portfolio baselines.

[26] Figure from Bill Monsen's CDRC Testimony from CA PUC A08-06-001, November 2008, p42.

Section 4. Conclusions

Baselines matter. They are a critically important foundation for good DR resource design, and as programs continue to expand, the necessity for clear, reliable measurement and verification standards becomes an even bigger contributor to grid reliability.

Previous studies have already helped inform demand response programs and they will continue to do so. Additionally, other novel designs already in practice are worth further study. The rolling average baseline, for example, has been used by ISO-NE for many years, and it would be useful to compare that method to High X of Y baselines.

Customers with volatile loads continue to be a challenge, and further analysis to develop and identify baselines that better manage these customers is needed.

Given what we currently know about baselines, as well as our experience with a wide range of DR programs, EnerNOC has laid out our findings in this whitepaper regarding what baselines are more or less appropriate than others. We conclude that—for most peak load management applications—High X of Y baselines with day-of adjustments represent the best balance of accuracy, simplicity, and integrity. While M&V will almost always require some nuanced considerations around system needs and customer engagement, these three pillars of good baseline design should be the starting point when drafting program rules.

Going beyond this paper, we will continue to work with other aggregators, utilities, grid operators, and regulators to improve baseline methodologies as technology, data availability, and other factors create new opportunities for accurate, simple-to-execute, and high-integrity measurement.

ENERNOC

ENERNOC

EnerNOC, Inc. Headquarters

101 Federal Street
Suite 1100
Boston, MA 02110
Office: 617.224.9900
Fax: 617.224.9910

EnerNOC Ltd. Headquarters

11-1155 North Service
Road West
Oakville, Ontario L6M 3E3
Office: 416.461.4678
Fax: 289.291.4001

EnerNOC UK Limited Headquarters

Alder Castle, 4th Floor
10 Noble Street
London EC2V 7JX
Office: 0800.520.0303
Fax: (0)800.520.0192

www.enernoc.com

For a full list of offices, please visit:
www.enernoc.com/about/contact.php

EnerNOC, Inc. is headquartered in Boston, MA, United States, with wholly-owned subsidiaries in Canada (EnerNOC Ltd.) and the United Kingdom (EnerNOC UK Limited). EnerNOC UK Limited is a company incorporated in England and Wales with company number 06937931, VAT number GB980145422 and whose registered office is located at Alder Castle, 4th Floor, 10 Noble Street, London EC2V 7JX. A list of directors is available for inspection at our offices.