Integration of Billing and Metering Data

Final Report

By
Xenergy, Inc.
1930 13th Street, Suite C
Boulder, Colorado
303/442-1990

Contributing Authors:
Dan Violette, Xenergy
Rich Barnes, Xenergy

Prepared for
The California Demand Side Management Advisory Committee:
The Subcommittee on Modeling Standards for End Use Consumption
And Load Impact Models

December 1994

INTEGRATION OF BILLING AND METERING DATA

1.1 INTRODUCTION

This paper describes five methods that may be used to integrate data from different sources or results from previous studies in an evaluation of gross energy impacts. The primary focus of this paper is the integration of billing data and metering data. Some of the methods discussed, however, also are applicable in integrating results from other evaluation methods such as engineering analysis.

Typically, integration in applied in one of the following situations:

1. Combining results from two or more independent studies;
2. Incorporating the results of previous studies into a current analysis; or
3. Using data from two or more sources in a single approach.

An underlying premise of integration is that previous research may be just as useful as present research. Integration is based on the assumption that the most recent study is not necessarily the best study, and that present research should build upon prior research. Thus, rather than following the most common evaluation practice of relying primarily, if not exclusively, on the results of the most recent study, integration provides a method to incorporate results from previous research in a current analysis.

Integration is used for two primary reasons: to reduce data collection costs, and thus evaluation costs, and to reduce bias. Integrating data from two sources provides a means to reduce overall data collection costs by leveraging higher cost data with larger samples of lower cost data. Integration also is used to reduce bias associated with specific approaches by combining the results of two or more approaches with different biases.

Many statistical methods make use of prior site-specific engineering estimates of savings. The accuracy of the prior engineering estimate is a key factor in determining the required sample size for a statistical analysis. Low cost methods involving on-site surveys or monitoring can be used to improve the accuracy of the engineering estimates before statistical analysis is conducted. By improving the accuracy of the engineering estimates, the required sample size for the statistical analysis may be reduced, thereby reducing the cost of the evaluation.

Integration may reduce bias by combining results of different evaluation approaches. Each evaluation approach contains some errors that do not diminish as the sample size increases. These types of systematic errors are referred to as bias. The potential biases of each approach differ as the measurement error, non-response, and specification error differ. By combining the results of two or more approaches that tend to have different potential biases, integration provides a potential means of reducing some bias.

Integration commonly is used to combine billing data with metering data. Whereas billing data almost always appears as consumption data, metering data may appear in three different forms: as end-use energy consumption data, as end-use metering and class load research studies, and as estimates of key engineering assumptions such as operation hours or connected loads. Metering gathers information on energy consumption or other key factors that determine how much energy is consumed. Metered energy consumption data is usually high resolution data with frequent observations covering short time periods. The metered consumption data tends to be at an end-use or equipment specific level.

End-use energy consumption monitoring can be conducted on a sample of participants. The monitored data can be used to derive an estimate of savings. Estimates of savings that are derived from the monitoring data usually are compared to tracking system estimates of savings. Either ratio estimation or regression analysis is used to leverage the tracking system estimates.

Energy consumption metering data often exists from end-use metering and class load research studies. This information is useful in calibrating some types of engineering models. If the engineering model produces estimates of end-use loads, it can be tested on the load research sites by comparing the engineering estimates to the metered data. Regression analysis is useful in conducting the comparison.

Metering also can provide estimates of key engineering assumptions such as operation hours or connected loads. This information is used to improve the accuracy of engineering models by improving the accuracy of the key assumptions. Detailed system performance monitoring of HVAC or process loads also may be useful in increasing the accuracy of the engineering assumptions.

1.2 INTEGRATION APPROACHES

The five general types of integration approaches are:

1. Weighted Average;
2. Bayesian Analysis;
3. CEM (Calibrated Engineering Model)/SAE (Statistically Adjusted Engineering Model) combination;
4. General Linear Model; and
5. Double Ratio Estimation with Billing Analysis.

Within each approach, a considerable amount of flexibility exists concerning sample sizes and the manner in which metering data is incorporated. The five approaches are described below. The typical applications for each integration approach are provided after the descriptions.

1.2.1 Weighted Average Integration

Weighted Average Integration combines the results of two or more studies by taking a weighted average of the results. The weights assigned to each study result are proportional to the inverse of the precision of each result. Thus, the lower the standard error of a given estimate, the more weight that estimate is given in determining the average.

The weighted average approach is very simple to apply. It can be used to combine any two independent estimates of the same variable. When the results from two different types of studies are averaged there may be an overall reduction in bias.

Weighted average integration should only be used when the two studies estimate the same specific variable. Combining the results from a current study with previous results may not be appropriate if the previous results were based on dissimilar programs or if program participants display very different characteristics.

Although the weighted average approach does combine the results of two independent studies, and thus may help to reduce bias, it does not provide a means to reduce evaluation costs. The weighted average approach does not take advantage of data leveraging, and does not provide a means of assigning evaluation costs to two or more different data sources in an optimal manner. With data leveraging, low cost estimation methods can be integrated with high cost methods to reduce the number of high cost data points needed to reach the required precision levels. The methods described below, unlike weighted averaging, do take advantage of data leveraging.

1.2.2 Bayesian Analysis

Bayesian analysis provides a method to incorporate previous findings into a current analysis. The Bayesian method allows for a cumulative consideration of what has been learned from various studies over time. Samples of new observations are analyzed in light of prior savings estimates. The sampled data is used to modify prior information on the probability distribution of savings. A “posterior” distribution is developed by considering the likelihood of the sample data occurring given on the prior probability distribution.

A Bayesian integration of metering data and billing data can occur starting from either the metering or billing data. The mean and variance of the probability distribution for the estimate of interest can be produced from an analysis of metered data. The prior information on the probability distribution is used in a Bayesian regression analysis of billing data. On the other hand, a billing analysis can be used to produce the prior estimates that are incorporated in a Bayesian analysis of metered data.

The prior information used in a Bayesian analysis should be based on previous studies and not set arbitrarily. Underestimating the variance of the probability distribution will bias the likelihood calculations of the sampled data and thus give more weight to the value of the prior estimate in determining the new revised estimate. Overestimating the variance of the prior distribution will tend to reduce the importance of the prior estimate in deriving the new revised estimate.

1.2.3 CEM/SAE Approach

The CEM/SAE approach is a two-stage analysis method. In the first stage, a CEM analysis is conducted in which engineering estimates of savings are produced or enhanced through the use of monitoring data and on-site observations. The second stage involves the use of a statistical regression model with prior engineering estimates of savings as an explanatory variable. The type of regression analysis often is referred to as a Statistically Adjusted Engineering (SAE) model.

In stage one, metering data is used to validate the assumptions used in the tracking system algorithms or to calibrate the tracking system estimates. The goal of the first stage is to develop a set of enhanced engineering estimates of savings that have a higher correlation with the "actual savings" than the original tracking system estimates. Applying a single ratio adjustment to all tracking system estimates does not increase the correlation and will have no effect on the SAE model results. A higher correlation to actual savings can only be obtained by applying different adjustments to different groups of customers.

The second stage involves using a SAE model in which the enhanced estimates from the first phase are used as priors. If the enhanced engineering estimates are more "accurate" than the original engineering estimates, then the standard error of the realization parameters will be lower for a given sample size.

The CEM analysis should be used to develop a set of enhanced engineering estimates for all participants. The SAE model is estimated using a sample of participants. The estimated realization parameters are applied to the population of the enhanced prior estimates. Developing enhanced engineering estimates just for the cases used in the SAE model estimation will not improve the precision over a SAE analysis that uses the original tracking estimates as priors.

The CEM/SAE approach can provide information that is useful in allocating the optimal amount of metering and billing observations. The value of the CEM stage can be assessed by comparing the relative precision of the realization rates with the enhanced engineering priors and the original engineering priors. The value of the CEM stage can be expressed in terms of the number of SAE sample points that can be reduced when enhanced prior estimates are used.

The metering data to be used in a CEM analysis may already exist from an end-use metering study or a class load study. Engineering models can be used to derive estimates of end-use loads or hourly whole building loads for the buildings in the load research sample. The engineering models are calibrated by comparing the engineering estimates to the metered data.

1.2.4 General Linear Model (GLM)

In a statistical analysis of billing data, considerable effort is required to explain the energy use of end-uses not affected by a program. The use of billing data requires more variables to be observed and included in the model specification than metering data. The potential for specification error and observation error would be reduced if the billing data only contained the energy consumption of the affected equipment. Metering provides a means to isolate the consumption of the affected equipment and reduces the data requirements of the analysis.

Metering data, though, introduces a few problems of its own. Non-response error is likely to be an issue in obtaining representative metering samples. End-use metering also may introduce some additional measurement error that does not exist in billing data. Metering data also is relatively expensive to gather, creating a situation in which only a small, potentially biased sample is possible.

A better research design may involve using observations from billing analysis and metered observations to estimate the parameters of a statistical model. The values of the independent variables for each observation in the model depend on the scope of the metered consumption data for each observation. On an observation by observation basis, all independent variables that only affect a given end-use can be set to zero if the monitoring data does not include consumption data for the given end-use.

The variance of the regression error term is expected to be smaller for metered data than whole building billing data. Standard regression estimation methods such as OLS (Ordinary Least Squares) are based on the assumption that the variance of the error term is constant across all observations in the model. In cases where this assumption of equal variance does not hold, a weighted least squares (WLS) estimation method is required.

Often, information can be obtained on the relative variances of the error terms for both sets of observations. A simple way to get this information is to estimate separate regression models for each source of data. The root mean squared error (RMSE) obtained for the separate regression analyses are used to develop weights for a weighted least squares estimation of the combined data. The value of the weight for the metered observations is set equal to the inverse of the RMSE obtained from a regression model that was estimated using only metered data. The weight for the billing observations is derived as the inverse of the RMSE obtained from a regression analysis of billing observations.

The integrating of metering data and billing data using GLM is not limited to a consumption change model. A conditional demand model using a SAE format often is used to calibrate engineering models to billing data. If metered end-use data is available, it can be used in the estimation process along with the billing data.

1.2.5 Double Ratio Estimation with Billing Analysis

In a double ratio estimation approach, two sets of ratio estimators or realization rates are developed. The first ratio is obtained by comparing tracking system estimates to an estimate of savings developed by using a relatively low cost approach. The second ratio is produced by comparing the results from the lower cost approach to results from a more accurate, more expensive approach that usually involves metering. The overall realization rate is simply the product of the two ratios.

The sample size used to develop the first ratio must be sufficiently larger than the sample used for the second ratio. Often the smaller, high cost sample is nested in the larger, low cost sample.

Since billing analysis is a relatively low cost approach, it is an excellent candidate for developing the first ratio. If a model with engineering priors is used, the parameter estimates are used to directly determine the value of the first ratio. If a model without priors is used, the value of the ratio estimator is the sum of the predicted savings from the billing analysis divided by the sum of the tracking system estimates for the customers who were included in the billing analysis.

The value of the second ratio is determined by comparing estimated savings from metering to the estimated savings from the billing analysis. The comparison is done just for the cases in which metering data was collected.

The use of billing analysis provides several advantages over a single ratio approach of metering data to tracking data. The billing analysis can reduce the number of monitoring observations that are required. The billing analysis also can reduce potential non-response error that often occurs with small metering samples.

The required sample size to produce a ratio estimator at a given precision depends on the correlation of the metered estimate with the tracking estimate. If the correlation between the billing analysis estimates and the metered estimates is higher than the correlation between the tracking estimates and the metered estimates, then a smaller sample of metering observations will be required. If the increased cost of the billing analysis is less than the reduced cost of conducting less metering, then the double ratio estimator will produce the same precision level of a single ratio estimator but at a lower cost.

1.3 APPLICABILITY OF APPROACHES

Selecting the appropriate integration approach depends on the objectives of the integration and on the types of information available. The weighted average approach is used to combine the results of two or more completed studies. The Bayesian approach is used to incorporate previous results in a current analysis of data. The other three approaches provide methods to leverage accurate, high cost observations with lower cost observations. Table 1 summarizes the most common situations where each method applies.

Table 1-1

Applicability of Approaches

Weighted Average	The approach is applicable when two or more independent studies exist and provide point estimates and precision estimates for the same variable. A key issue is ensuring that both studies are actually estimating the same variable. The savings of a measure can change over time as the technology changes or as it is adopted by different market segments. The most appropriate situation for using weighted average integration is the case where two approaches that rely on different data sources and estimation methods are used to estimate the same variable.
Bayesian	Bayesian analysis is applicable any time prior information is available on the parameter of interest. One must take into consideration that the actual savings of a measure could be changing over time. The estimated variance of the prior probability distribution should take into account all sources of error. The variance estimate can also be increased to account for the situation where savings change over time.
CEM/SAE	Using CEM to improve the accuracy of the prior estimates in a CDA model will only be cost-effective in a limited set of situations. First, the accuracy of the original tracking estimates must be low enough so that significant improvement is possible. The condition tends to occur only when tracking estimates are not site-specific. Second, the cost of obtaining monitoring data must be very low. This cost limitation implies the use of run-time loggers and spot load measurements. Existing load research data also provides a low cost source of consumption data that can be used to calibrate an engineering model.
GLM	End-use metered data can be included in a classic CDA analysis of billing data. The use of metered data will produce UEC/EUI estimates that tend to be more precise. If a time-series or pre/post CDA model is desired, the time series observations of the metered data must be defined in a manner consistent with how the billing data observations are defined.
Double Ratio	Billing analysis provides a method to improve the accuracy of the tracking estimates before being compared to “metered” observations of savings. The use of billing analysis can be cost-effective if the accuracy of the original tracking estimates is low and the cost of a metered observation is high.

1.4 ASSESSMENT OF APPROACHES

The five integration methods may be assessed in terms of five attributes. The five attributes considered were:

- UEC / EUI estimates;
- Data requirements;
- Costs;
- Errors; and
- Robustness.

The attributes were selected to be consistent with the evaluation of statistical and engineering models that was recently performed for CADMAC. Please refer to the June 1994 report, "An Evaluation of Statistical and Engineering Models for Estimating Gross Energy Impacts, Final Report," for definitions of the five noted attributes.

The characterization of each approach is from a general perspective. The cost, errors, robustness, and data requirements can vary significantly within a given integration approach. The conclusions provided in Table 1-2 through 1-6 only represent the tendencies of the approaches. The attributes of a specific integration approach may be very different from the general tendencies noted below.

Table 1-2

Evaluation Matrix for Integration Approaches: UEC/EUI Estimates

Weighted Average	UEC / EUI estimates can be developed if at least one of the contributing studies provide these estimates.
Bayesian	UEC / EUI estimates can be produced if a classic CDA approach is used. Bayesian priors of UEC / EUI can also be incorporated into the analysis.
CEM / SAE	The ability to derive UEC / EUI estimates depends on the type of statistical model used. Classic CDA models will produce UEC / EUI estimates. Non-classic CDA models will not produce estimates of UEC / EUI
GLM	
Double ratio	Generally, UEC / EUI estimates are not produced using this approach.

Table 1-3

Evaluation Matrix for Integration Approaches: Input Data Requirements

Weighted Average	Estimated savings from two or more independent studies are required. Precision estimates are used to determine weight given to each estimate.
Bayesian	Data requirements are: a prior estimate of savings obtained from one or more previous studies, estimated standard error of the prior estimate, and a sample of either billing or metering observations with explanatory information obtained from utility's data bases or surveys.
CEM/SAE	Billing data, tracking estimates of savings, and survey data requirements depend on the type of SAE analysis that is chosen. Low-cost monitoring data and on-site observations are needed from a sample of participants. Monitoring data usually consists of connected loads and run-time hours.
GLM	Both end-use/equipment metering and billing data are used in GLM. The metering data and billing data must both meet the requirements of the statistical models that have been selected. Survey data requirements also depend on the statistical model used. Usually, less survey data is required for metered sites.
Double ratio	Tracking estimates of savings and realization rates from a statistical analysis of billing data are needed. The realization rates must vary across two or more market segments. Estimates of savings derived from metering for a sample of participants are also needed.

Table 1-4

Evaluation Matrix for Integration Approaches: Costs

Weighted Average	A simple and very low cost method to integrate the results of two or more existing studies.
Bayesian	The cost of a Bayesian regression analysis is only slightly higher than the cost of a classical regression analysis.
CEM/SAE	Monitoring costs tend to be lower than the GLM or double ratio approaches. Analysis costs are higher since both a SAE and CEM analysis are needed. A SAE model used by itself will tend to be a more cost-effective method unless the CEM process can significantly increase the accuracy of the tracking estimates.
GLM	Although the cost continues to drop, time series metering of end-use kWh consumption still tends to be relatively expensive. Some additional analysis costs are needed to determine the appropriate estimation weights.
Double Ratio with Billing Analysis	End-use consumption monitoring is expensive especially if pre-retrofit monitoring is required. Analysis costs are similar to the GLM method although the steps are somewhat different. The cost of this approach tends to be lower than the more common double ratio approach where CEM is used to produce the first phase estimate.

Table 1-5

Evaluation Matrix for Integration Approaches: Errors

Weighted Average	The sampling error is less than the sampling error of any of the individual studies. Non-sampling errors will tend to persist although canceling of bias is possible.
Bayesian	A Bayesian analysis will produce estimates of precision that are smaller than a classical analysis. The increase in precision is due to the use of prior information. Some canceling of bias can occur when different analysis methods are integrated.
CEM/SAE	The use of CEM to enhance the tracking estimates will reduce sampling error. A similar reduction in sampling error could also occur from increasing the sample size used in the SAE model. Since the tracking estimate is an independent variable that is observed with error, the CEM process can also reduce the "observation" error of the prior estimate.
GLM	The uses of metered data will tend to reduce specification error and observation errors of independent variables. Measurement errors and non-response errors can increase when metering observations are used.
Double ratio	The use of billing analysis will reduce sampling error and non-response error. The use of metering data as the "gold standard" will increase the potential for significant measurement error.

Table 1-6

Evaluation Matrix for Integration Approaches: Robustness

Weighted Average	The robustness of the individual studies will directly determine the robustness of the integrated "Weighted Average" estimate. The new estimate will tend to be more robust than the individual estimate.
Bayesian	The robustness of a Bayesian estimate is largely dependent on the robustness of the prior estimate. Incorporating previous results as prior estimates into new studies will tend to produce a more stable estimate over time. Robustness is further enhanced by applying the results from different approaches over time.
CEM/SAE	The robustness is generally the same as any regression model approach with engineering priors.
GLM	
Double ratio	Key assumptions involve the representativeness of the metering sample. Random samples are not easy to obtain when end-use metering is involved. Another important assumption involves the degree that measurement error exists in the monitoring data.