ERRORS IN VARIABLES: A CLOSE ENCOUNTER OF THE THIRD KIND

Richard Ridge, Ridge & Associates

Introduction

Both the evaluations of PG&E's 1992-93 Commercial New Construction Program (CNCP) and the 1994 Commercial HVAC Program (CHP) attempted to employ an SAE approach using engineering information collected in great detail at great cost. This approach appeared to work reasonably well in the CNCP where the estimated realization rate was .84. However, in the case of the CHP, the realization rates consistently hovered around .35 which, given our confidence in the quality of the engineering priors, seemed implausibly low. Eventually we were forced to opt for a dummy variable approach that produced a realization rate of .92. What follows is an attempt to explain the performance of the SAE models in the evaluation of these two programs.

This article will provide a review of the errors in variables problem and review some of the literature on the effects of this problem in DSM evaluation. It will then describe how econometric theory and recent Monte Carlo simulation studies were confirmed in a recent evaluation of PG&E's CNCP and the 1994 CHP, and suggest steps that can be taken to minimize this problem.

Measurement Error

The error in variables problem is well understood in the econometric literature. However, it is only recently that the effects of errors in variables have been investigated in the context of SAE models, a common modeling approach used in the evaluation of DSM programs.

It was during the 1980s that dissatisfaction grew with the use of dummy variables in evaluating DSM programs. This concern prompted the use of what appeared to be a more refined measure of the impact of installed energy efficient measures. This more refined measure, it was argued, should increase the precision of estimates of energy savings. The dummy variables were replaced in the regression equation by the expected annual or monthly savings, often referred to as an engineering prior. This prior could be entered into a regression model instead of the traditional dummy variable along with other independent variables such as weather and other changes in the house or building that were expected to affect kWh consumption. Equation 1 is an example of the general form of such a model.

$$E_{b,t} = \alpha + \sum \lambda \ X_{b,t} + \beta \ ENG_{b,t} + \varepsilon_{b,t} \qquad (1)$$

where

$E_{b,t}$ = metered energy consumption of building b at time t

α = model intercept

β = estimated adjustment coefficient for the engineering prior

λ = vector of estimated coefficients for explanatory variables

$X_{b,t}$ = a vector of site- and time-specific explanatory variables

ENG = engineering estimate of savings

$\varepsilon_{b,t}$ = captures the kWh reduction not explained by the model

The estimated coefficient for this engineering prior, β, often referred to as a realization rate, was an indication of the percent of the expected savings that were realized. These priors were most often based on simplified engineering algorithms that were in many cases prepared by utility staff responsible for implementing the DSM programs. Utilities usually estimate the expected kWh and kW savings associated with the installation of each energy efficiency measure, e.g., efficient air conditioner, efficient lights, insulation. These estimates along with other important customer information are maintained in program tracking databases (PTD). However, one of our major concerns was that the engineering priors contained tracking system priors may contained error. Of course, there are various types of error.

Random and Non-Random Error

There are two basic kinds of errors that affect empirical measurements*: random error* and *non-random or systematic error*. Random is the term used to designate all of those chance factors that confound the measurement of any phenomenon. The amount of random error is inversely related to the degree of reliability (precision) of the measurement instrument. That is, a highly reliable indicator is one that leads to consistent results on repeated measurements because it does not fluctuate greatly due to random error. The effects of random error are totally unsystematic in character. An engineering prior that contains random error is one that, in repeated measurements, sometimes overestimates the savings while at other times underestimates the savings. With respect to the estimation of HVAC savings, recent research suggests that some of this random error is probably due to the unreliability of estimates of operating hours (Sonnenblick and Eto, 1995).

The second type of error that affects empirical measurements is non-random or systematic error. Unlike random error, non-random error has a systematic biasing effect on measuring instruments. Thus, an engineering prior that contains non-random error is one that, in repeated measurements, always results in either underestimates or overestimates of HVAC savings. Non-random error is very much related to the concept of validity (accuracy) which is defined as the net difference between the obtained measurement and the true value. Just as reliability is inversely related to the amount of random error, so validity depends on the extent of non-random error.

In any given study, it may be the case that some of the variables being measured cannot be measured accurately, either because of data collection difficulties or because they are inherently difficult to measure. Random errors in measuring the dependent variables are incorporated in the disturbance term and their existence causes no problems. However, when the random errors are in the independent variables, the problems become quite serious, resulting in biased estimates. To see why this is the case, consider the following. Assume that:

$$x_i^* = x_i + v_i \qquad (2) \qquad (2)$$

where x_i is the true value
x_i^* is the observed value.

The true regression model is

$$y_i = \beta x_i + \varepsilon_i \qquad (3) \qquad (3)$$

Recognizing that $x_i = x_i^* - v_i$, the actual regression run is

$$y_i = \beta x_i^* + (\varepsilon_i - \beta v_i) = \beta x_i^* + \varepsilon_i^*) \quad (4) \qquad (4)$$

To say that x is measured with error is to say that it is not fixed in repeated sampling. Instead, the values are generated by a random process as reflected by v_i in Eq. 4. In other words, the “observed” independent variable is a random variable, called a stochastic regressor, that is not independent of the disturbance term. That is, the error ε^* and the variable x^* are correlated, i.e., have a nonzero covariance. Another way of looking at this problem is to refer to v_i in Eq. 2. The greater the variance of v_i, the greater the variable x_i^* resembles a random variable and the greater the correlation with ε^*. This correlation violates a fundamental assumption of OLS leading to a biased estimate of β.

In the case of a single explanatory variable, errors in measuring the variable cause the coefficient to be biased downward. Also, in the case where there is more than one independent variable in the model, one of which we are certain was measured with a fair amount of error, there is also bias in the other coefficients, although the direction is unknown. Finally, if two or more variables are measured with error, the situation is much more complex and one cannot determine the magnitude or the direction of the bias in any of the coefficients. (Johnson et al., 1987),

The case of systematic error is not a problem. If the engineering prior in a utility’s program tracking database is in every case greater than 10% of the true value, then this can be viewed simply as a scaled variable. When a variable is scaled up or down by a given factor, a, the coefficient is changed by the factor 1/a, the inverse of the factor used to scale the independent variable. The intercept remains unchanged. The important point here is that the estimated coefficient is changed to reflect the change in the scale of the dependent variable but it remains an *unbiased* estimate. (Johnson et al., 1987).

Because SAE models are well suited to estimate the amount of systematic error, such error presents no problem. For example, a realization rate of 80% indicates that a utility tended to overestimate savings systematically by 20%. If there is little random error in this utility estimate, then this estimate of 80% is unbiased. On the other hand, if there is a fair amount of random error in the measurement of the engineering prior, then the estimate of 80% is biased.

A Shared Concern

Over the last several years, the quality of engineering priors residing in utility program tracking systems has been called into question. In 1992, SCE (1993) estimated gross energy and demand impacts for participants in its 1990 Commercial Energy Management Hardware Rebate Program. DOE2 analyses were conducted for a total of 25 measures in the HVAC, Lighting, and Other end uses. Focusing on the HVAC end use, 73 buildings were examined in which package air conditioners were installed. The results of the sophisticated DOE2 analyses (ex post) were compared to the pre-installation (ex ante) estimates provided by Edison’s DSM program staff using simplified engineering algorithms and entered into Edison’s program tracking database. The errors in estimated kWh impacts were significant and appear to be primarily random rather than systematic. The simple correlation between coefficient the ex post estimates of kWh savings and the ex ante estimates was only .19.

Another example is the evaluation of PG&E’s Commercial HVAC Program. For this evaluation, estimates produced by the sophisticated DOE2 analyses (ex post) were compared to the pre-installation (ex ante) estimates provided by PG&E’s DSM program staff using simplified engineering algorithms and entered into PG&E’s program tracking database. The correlation between the DOE2-based estimates and those contained in the program database was .51. While this correlation is higher than in the SCE case, there remains a fair amount of random error.

Similar results at other utilities prompted Violette (1993) to study this problem using a using a propagation of error (POE) model. Using the POE model, Violette shows how one can examine the uncertainty surrounding each of the inputs contributing to the estimation of savings as well as its uncertainty. He points out that uncertainty is comprised of two components, systematic bias and random error. The former can be minimized using by "careful attention to measurement protocols and algorithm selection." The second can be addressed through the use of the POE model by which one can develop a cost-effective data collection strategy by examining the tradeoff between an increase in the cost of collecting more accurate inputs used to estimate savings and the resulting increase in the accuracy of the savings. One of the results of his analysis was that the operating hour variable contributes more uncertainty to the savings estimates than other variables.

Violette concludes that there are now two basic choices facing any evaluation analyst:

1. …they can improve the estimates of DSM program savings by developing better statistical models of facility energy use (or change in use) using more sophisticated methods and larger sample sizes
2. …the precision of the estimates of DSM program savings can be increased by improving the site-specific engineering estimates and, in particular, by improving the ability of the engineering estimates to explain site-to-site variation in impacts. (p. 656)

Vine et al. (1995) also expressed concern over the uncertainty surrounding savings estimates. Included in their paper is an attempt to clarify some important concepts. They provide a number of recommendations for reducing uncertainty. Two of these recommendations concern measurement and evaluation.

- Prepare guidelines for achieving cost-effective accuracy levels (similar to MDPU [Massachusetts Department of Public Utilities] decision). Consider how to obtain a given reduction in uncertainty in the most cost-effective manner.

- Prepare guidelines for reducing bias, so that key factors are accounted for (e.g., program spillover).

Sonnenblick and Eto (1994), like Violette (1993), also used a POE approach that was implemented via Monte Carlo simulation methods to explore uncertainty of savings estimates as a function of uncertainty surrounding key inputs to the savings calculation. Like Violette (1993), they found that operating hours are responsible for most of the uncertainty in the realization rate. They go on to add:

- Because the precision and bias of tracking database and site inspection estimates of savings seem to vary considerably, and because an evaluator, absent additional evaluation information, has no means of estimating the accuracy and precision of their tracking database estimate, it is dubious to rely upon tracking database estimates of savings alone.

More relevant for this present article, is their excellent and thorough analysis of the biasing effects of random error in SAE models. Using Monte Carlo techniques, they created 500 simulated commercial buildings each with two years worth of billing data. They created several versions of the engineering prior. In creating these versions, they were guided by errors observed in actual utility evaluations. One prior mimicked the error found in utility program tracking databases; one was somewhat less accurate and was based on site inspections in which auditors verified the existence and operation of the measured and adjust tracking database estimates based on interviews with customers; the third was the least accurate and mimicked the error observed in actual utility program tracking databases. Their model also included other variables such as building size, weather, and annual hours of operation.

Focusing first on priors in which there was no systematic error, they found that the greater the random error surrounding an engineering prior used in an SAE model, the greater the downward bias. In cases where there were less than perfectly accurate (containing random error) engineering priors, dummy variable models outperformed SAE models. In cases in which the prior was perfectly accurate (no random error), an unbiased estimate of the realization rate resulted. More accurate bottom-up (e.g., metering) are needed before inclusion in the regression models can improve savings estimates. Next, they focused on two situations in which there is random variation around a prior that contains systematic error and a prior that contains no systematic error. In situations in which the amount of random error was the same for both priors, they found that the former model outperformed the latter model. When the random error around both priors was reduced to less than 10%, both models produced estimates of realization rates near 1.

They concluded that the SAE method " . . . has dubious value unless the tracking system estimate used in the regression is very precise and reasonable unbiased." They also examined the costs of collecting data to support an SAE model and found that, if site inspection data has already been compiled, *three fold increases* in evaluation spending are required to increase model accuracy another 20% to 30%.

From all of these analyses described in this section, three major points emerged. First, the quality of savings estimates produced by simplified engineering algorithms is very likely poor. Second, the magnitude of the random error component appears to be significant. Third, the estimates can be improved using higher quality data and more sophisticated models *but* only at great cost.

The remainder of this article will attempt to underscore the problem of errors in variables in SAE models via a much less complex Monte Carlo simulation and discuss how the results of the simulation studies were verified in a recently completed evaluation of PG&E's 1992-93 Commercial New Construction Program and PG&E's 1994 Commercial HVAC Program. Finally, recommendations will be made with respect to the future use of engineering analysis in evaluating DSM programs.

Simulation Exercise

A simulation exercise was performed to illustrate the effects of systematic and random error on regression coefficients. The point here is not to replicate the work done by Sonnenblick and Eto (1995) but only to underscore in a considerably less complex analysis the problem of random errors in the engineering priors in SAE models.

A simple Monte Carlo simulation was conducted in which 400 program participant cases were created in order to illustrate the effects of measurement error. Each case was assigned 36 months of baseline kWh consumption by drawing from a random normal distribution from 10,000 kWh to 50,000 kWh per month. Next a month, between the 13th and 24th month, was randomly selected from a uniform distribution as the month when the efficient equipment was installed. At the month of installation, each participant's monthly consumption was reduced by 10% of the baseline consumption. Next, four engineering priors were created.

1. The first was set equal to 10% of the each participant's baseline consumption. This is clearly the most accurate prior with an expected β of 1.0.
2. The second was set equal to 90% of each participant's baseline consumption. This prior contains only systematic error with an expected β of 1.11 (1\.9).
3. The third was set equal to 110% of each participant's baseline consumption. This prior contains only systematic error with an expected β of .9 (1\1.1).
4. Other priors were created that contained varying amounts of random error on either side of the true value.
5. The fifth was simply a dummy variable, coded as a 0 before installation and a 1 beginning in the month of the installation.

The SAE model was estimated using SAS's PROC GLM procedure and contained only two variables, a customer specific intercept that captured differences in base usage across all customers and the installation variable. The customer-specific intercept was allowed for by using PROC GLM's ABSORB option.

$$E_{i,t} = \alpha_i + \beta ENG_{i,t} + \varepsilon_i \qquad (5)$$

where
$E_{b,t}$ = energy for customer i at time t
α_i = customer-specific intercept
β = estimated adjustment coefficient for the engineering prior
ENG = engineering estimate of savings
$\varepsilon_{i,t}$ = captures the kWh reduction not explained by the model for the ith customer at time t.

Note that this simulation is much simpler than Sonnenblick and Eto (1995). First, there is only one independent variable. Second, the monthly consumption for each customer does not vary except, of course, for the one 10% change following the installation of the efficient equipment. This means that there is no noise in the monthly consumption data. This simplification is done in order to focus on the biasing effects of random error in priors while holding all other variables constant.

In Table 1, the results of these simulations are presented. The first three results are for the completely accurate prior and the two priors with 10% systematic error. The next eight models used priors that had varying amounts of random error. The final model employed the dummy variable. Reported for each run is the beta (the realization rate) and the magnitude of the error as a percent of the true value.

As one can see, realization rate for the accurately measured prior is 1, as was expected. For those priors measured with a 10% systematic bias, the realization rates are also what one would expect. For those priors containing random error, the realization rates are biased downward as the random error component becomes greater.

This simulation suggests that the dummy variable approach is superior in situations where the engineering prior contains random error that exceeds approximately 15%.

This simulation suggests that the dummy variable approach is superior in situations where the engineering prior contains random error that exceeds approximately 15%.

Close Encounters of the Third Kind

Before going any further, let's clarify a few key terms. Any real-life encounter with something unusual or unexpected (i.e., extra terrestrials, ghosts, honest politicians, measurement error) has been referred to an encounter of the third kind. Encounters of the fourth kind are when you are actually taken hostage by what you've encountered. The two evaluations reviewed in this section describe a situation in which the problem of measurement error was reduced to a minimum and one in which the problem of measurement error became an encounter of the third kind that almost became an encounter of the fourth kind.

Table 1. The Biasing Effects of Random Error in Engineering Priors

Variable	Realization Rate	Magnitude of Error As Percent of True Value
No Error	1.00	0
Systematic Error: 10% Overestimate	.91	+10
Systematic Error: 10% Underestimate	1.11	-10
Random Error #1	.99	+/-10
Random Error #2	.95	+/-20
Random Error #3	.89	+/-30
Random Error #4	.82	+/-40
Random Error #5	.76	+/-50
Random Error #6	.69	+/-60
Random Error #7	.64	+/-70
Random Error #8	.59	+/-80
Dummy	.98	N/A

The two evaluations reviewed in this section are: 1) PG&E's 1992-1993 Commercial New Construction Program, and 2) PG&E's 1994 Commercial HVAC Retrofit Program. These studies have been chosen because they illustrate several important points regarding measurement error and the engineering costs associated with using the SAE approach. The focus will be on estimation of gross impacts for the HVAC end use using SAE models that involve only program participants.

Evaluation of PG&E's 1992-1993 Commercial New Construction Program

This evaluation employed a very expensive, complex, and comprehensive engineering and statistical effort. However, in this paper, the focus will be on those elements that are germane to the issue of measurement error in the context of SAE models.

On-Site Surveys. This evaluation involved the collection of data for 171 jobs[1] at 150 program participant sites using on-site surveys. Of these 150 sites, 36 were associated with HVAC installations. These surveys focused on the area served by the control number[2] associated with the sampled job. The data collected at the level of the control number provided information needed for the statistical analysis of gross impacts. Collecting data at this level meant that one could examine those kWh data expected to change as a result of the installation rather than the kWh data associated with the entire site. This meant that the savings as a percent of the consumption (*the effect size or the single-to-noise ratio*) were larger thus increasing the power of the test. In some cases, the scope of the area surveys was defined by more than one control number.

Engineering Analysis. *While the engineering analysis for all 36 sites consisted of five stages, for the purposes of this article, only three will be reviewed.*

1. As-Built with Program Algorithm (Evaluation Method 1): Produced a revised estimate of program savings using the same algorithms used by the Program but incorporating as-built conditions observed in the on-site survey as well as data obtained through short-term end use metering for a sample of sites. For the HVAC end use, 7 sites were sampled for short-term metering in order to determine lighting schedules, HVAC operating hours, and motor utilization.
2. As-Built with Best Available Algorithm (Evaluation Method 2): Produced a revised estimate of program savings using the best available algorithm and incorporating as-built conditions observed in the on-site survey. A number of engineering models were used, including DOE2.1E simulation model. In some cases, the best available algorithm was identical to the program algorithm. The realization rate (new engineering estimate/original tracking system estimate) for cooling, using Method 2, was .86.
3. Evaluation Method 2 Enhanced by Summer Metering (Evaluation Method 3): For certain sites, the modeling of measure performance was substantially improved by obtaining end-use metering data during the summer of 1995. Multi-channel loggers were used to meter hourly cooling kWh for seven out of the 43 jobs (at the 36 sites) where condensing systems, efficient chillers, or efficient package units were installed. Power measurements were tailored to the specific equipment configuration at each site.

Next, a method was developed to enhance the estimates of annual kWh for the 29 sites that did not benefit from metering. First, for each of the seven sites the following model was estimated:

$$KWHSQFT_{HVAC} = \alpha + \beta_1 TEMP_h + \beta_2 HOUR + \beta_3 HOUR^2 + \beta_4 HOUR^3 \quad (6)$$

where
$KWHSQFT_{HVAC}$ = the hourly cooling kWh per square foot of measure-affected conditioned floor area

[1] A job is defined as a collection of measures described in a rebate application submitted by a commercial customer. In some cases, more than one job was completed at a customer site.

[2] When electrical service is established at a new location, a meter base is installed. PG&E assigns a permanent number to this meter base. Over time, one or more meters may be installed to measure electrical energy supplied through the meter base, but each of these meters is linked to the permanently assigned control number.

$TEMP_h$ = temperature at a given hour *h* of the day
HOUR = the hour of the day (1 though 24)

Each of the 29 remaining sites was then matched to one of the seven sites based on building type and HVAC system type. Each of the seven estimated models was then used to predict the hourly cooling kWh per square foot of measure-affected conditioned floor area for each of the sites matched to it.

Statistical Analysis. SAE models were then used to estimate the gross energy impacts. The models were estimated first in early June without the benefit of the summer end-use metering and then later in the fall with the benefit of the end-use metering. The realization rate for HVAC for the first estimation was .69. Once the engineering priors for the 43 cooling sites had been enhanced using the summer metering data, the realization rate rose to .84.

We suspected that the engineering priors for all 36 sites even without summer metering had little random error to begin with since they were based on very careful site-specific treatment. As a result, a fair amount of both the systematic and random error was very likely eliminated. It appears that the improvement in the realization rate was due to the correction at the 29 sites of systematic bias in the operating hour variable.

Evaluation of PG&E's 1994 Commercial HVAC Retrofit Program

In September of 1995, the evaluation of PG&E's 1994 Commercial HVAC Rebate Program was well underway. The research plan for the evaluation of PG&E's 1994 Commercial HVAC Rebate Program also called for a great deal of costly and sophisticated engineering analysis followed by statistical analyses designed to estimate both gross and net impacts.

Engineering Analysis. The purpose of the engineering analysis was to provide better engineering priors for use in SAE models and to provide a backup estimate of gross impacts if the statistical models were ill behaved. Our solution, consistent with the preliminary findings of Sonnenblick and Eto (1995), began with an effort to reduce the uncertainties in savings estimates through the use of detailed site inspections, metering, and DOE-2.1E analyses. A total of 139 commercial participants received on-site surveys in order improve the engineering-based estimates of the gross savings contained in PG&E's program tracking database.

It was expected that this effort would reduce both random *and* non-random error in the savings estimates *for the 139 sites*.

An approach was developed to leverage detailed information about one building by applying the information to other similar buildings, thus maximizing the number of sites that could be analyzed. Prior to the on-site survey, the 139 sites chosen for the engineering impact evaluation were divided into two groups, a cluster group of 60 sites and a matched pair group of 79 sites. Sites in the cluster group received a more intensive on-site survey and a DOE 2.1E analysis calibrated to monthly bills. The initial analysis of the cluster sites was first completed, and then information from that analysis supported the matched pair analysis. These analyses are described in more detail below.

On-site surveys for the 60 cluster sites involved collecting data to characterize the as-built and pre-measure capacity, efficiency, and quantity of the measure-affected equipment. Surveyors also collected data on the type of HVAC system, operating schedule, control settings and other performance parameters, as well as the operating schedule for internal loads in the conditioned spaces served by the affected HVAC system, the power density of internal loads in those spaces, and the building envelope characteristics (conditioned floor area, number of floors, percent glazing, and glazing type).

Once the surveys were completed, the 60 cluster sites were grouped into five sets according to key building characteristics. These groups were: (1) school, (2) retail, (3) hospital, (4) office with central A/C, and (5) office with packaged A/C.

After the clusters were defined, five calibration sites were selected to represent groups of the cluster analysis sites and three test sites that were used to assess the value of the clustering approach to energy savings estimation. Calibrated simulation models were prepared for each of these sites using the data collected from the on-site survey, along with billed 1994/95 gas and electric consumption and actual 1994/95 hourly weather for the closest NOAA station (supplemented with PG&E temperature data). In addition, for five of the calibration sites, the simulation model was calibrated to the short-term end-use metering data described.

A site-specific calibration plan was developed for each calibration and test site. Per the specifications of this plan, the model was calibrated for each calibration and test site against actual consumption (kW, kWh and therms) for the post-installation portion of the 1994 summer cooling season (July, August and September). Simulation inputs were prepared using survey data. Short-term end-use metering data from the early part of the summer of 1995 was also used to establish realistic internal load schedules and control logic for the HVAC system in the five calibration sites.

Once each of the calibration and test site post-period models were complete, each of them was used to estimate typical base and efficient post-period use (gas and electric) for the corresponding cluster of cluster analysis sites. As-built consumption for each site was calibrated to within 10% of billed kWh and 20% of kW for a calibration period in 1994. After calibration, the cluster model was rerun using typical weather for the pre-condition, as-built, and, when appropriate, Title 20 baseline cases. Gross savings were calculated by subtracting as-built consumption under typical weather conditions from pre-condition consumption.

The question is, in our attempt to leverage the information from the five calibration sites, did we manage to eliminate or at least diminish the random error. Recall that, in order to determine whether this was the case, a test site was chosen from three of the five clusters in order to assess the accuracy of the clustering approach. These test sites were calibrated twice: as though they were a calibration site (except for end use metering) and as though they were a typical cluster site. The resulting estimates of savings could then be compared to see if our attempt to leverage the information contained in the calibration sites was successful.

Recall that three test sites were modeled as both calibration and cluster sites to determine what effect the clustering process would have on the accuracy of the savings estimates. The three sites were deliberately chosen to rep represent a range of system types, building types, and HVAC measures. Consumption and savings estimates for the three sites are shown below in Table 2. The aggregated estimates showed very small differences: the sum of the as-built consumption for the three buildings, for instance, showed a difference of 2% between the two methods. Aggregate electric consumption savings estimates were within 1.2% of each other; aggregate gas savings were within 13.7% of each other. Overall, the total savings (both electric and gas combined) differed by 11.7%.These results confirmed that DOE 2.1E modeling using a clustering approach yielded savings estimates reasonably close to those generated by detailed, site-specific DOE 2.1E models. *However, it is more salient for the SAE modeling that the errors in the kWh savings estimates at each of these three sites are 35.6%, -34.4%, and 10% and they appear to be random*.

On-site survey data for the 79 matched-pair sites were similar to those for the cluster sites, although with *somewhat less detail* about the specifics of the HVAC system. Based on data about building type, size, envelope characteristics and HVAC system type, each matched-pair site was paired with an appropriate cluster site. Key parameters of the DOE 2.1E model for that cluster site, such as HVAC schedules, setpoints, and glazing percentages, were then modified to reflect the matched-pair as-built and pre-measure conditions. As with the cluster analysis sites, gross savings were calculated by subtracting as-built consumption under typical weather conditions from pre-condition consumption.

Table 2: Cluster Test Site Comparison

	TEST METHOD	CLUSTER METHOD	% DIFFERENCE
SITE 1: Office with packaged A/C units			
Total As-built Usage (kWh/year)	1,374,917	1,373,978	-0.07%
Electric Savings (kWh/year)	9,059	15,819	74.6%
Gas Savings (kWh/year)	100,994	133,449	32.1%
Total Savings (kWh/year)	110,053	149,268	35.6%
SITE 2: Office with chillers			
Total As-built Usage (kWh/year)	1,252,969	1,198,201	-4.3%
Electric Savings (kWh/year)	32,249	21,139	-34.4%
Gas Savings (kWh/year)	--	--	--
Total Savings (kWh/year)	32,249	21,139	-34.4%
SITE 3: School with absorption chillers			
Total As-built Usage (kWh/year)	782,363	770,643	-1.5%
Electric Savings (kWh/year)	139,395	145,884	4.6%
Gas Savings (kWh/year)	747,736	831,270	11.1%
Total Savings (kWh/year)	887,131	977,154	10.0%
TOTAL FOR ALL SITES			
Total As-built Usage (kWh/year)	3,410,249	3,342,822	-2.0%
Electric Savings (kWh/year)	180,703	182,842	1.2%
Gas Savings (kWh/year)	848,730	964,719	13.7%
Total Savings (kWh/year)	1,029,433	1,147,561	11.5%

It is important to point out that errors introduced in the cluster analysis were propagated to the 79 matched-pair sites. We suspect that additional errors were introduced since the matched-pair sites used less detailed data and relied more heavily on default values associated with such factor as their building type and climate zone.

Next, we attempted to leverage the improved data from the 139 sites to improve the priors in the remaining 311 sites that did not receive on-site inspections or DOE2 modeling. Recall that of these 139, 60 were cluster-analysis sites and 79 were matched-pair sites[3]. Our original expectation was that this effort would reduce both random *and* non-random error in the savings estimates for the 139 sites. However, recall that Table 1 clearly indicates that, while the total savings for all three cluster sites was within 11.5% of the estimate provided by a rigorous engineering analysis, a fair amount of random error was introduced for each of three sites. This suggests that, for the 52 cluster sites, significant random error was introduced. This in turn meant that the 79 matched pairs suffered from further propagation of these errors. We also suspect that little of the random error in the priors for these 79 sites was reduced since for these sites there was less reliance on site-specific data.

Without fully understanding the extent to which we had failed to reduce the random error, we proceeded to use the enhanced priors for the 139 sites to minimize the systematic and random error for all the remaining 311 sites. Note that we planned to estimate the gross model using all 450 sites (139 + 311). To do this, we used a single ratio approach which involves first calculating the ratio for each of the 139 sites of the enhanced engineering-based estimates of gross savings to the original PG&E engineering-based estimates of gross savings contained in the program tracking database. This ratio is in effect a realization rate. Next, each PG&E estimate for the 311 sites was then adjusted up or down by multiplying it by this ratio. This adjustment for the other 311 customers is in effect a prediction of what the enhanced estimates would have been had these other 311 customers also received on-site surveys and subsequent simulation analysis. We suspected that the original priors residing in the program tracking database contained a fair amount of random error.

However, simply adjusting these priors up or down using these ratios from the 139 sites no doubt did very little to mitigate the random error problem for the 311 sites. Why is this the case? First assume that the original variables contain random error. When these variables are scaled up or down by a given factor, a, the coefficient is changed by the factor 1/a, the inverse of the factor used to scale the independent variable. The intercept remains unchanged. The important point here is that the estimated coefficient is changed to reflect the change in the scale of the dependent variable but it remains a *biased* estimate since systematically scaling each variable does nothing to eliminate the random error in the original variable (Johnson et al., 1987).

Another important issue is the fact that the PG&E estimates of gross savings are annual rather than monthly. If one wishes to use a monthly model, then one must allocate the annual savings to months in a manner that recognizes any seasonal patterns. Aggregation of the measure estimates from the DOE 2.1E simulation analyses, by typical month, served as a useful method to allocate PG&E kWh savings estimates across months. Thus, for all sites, except those five sites that received careful calibration, error was introduced to the extent that any building did not match one of the five buildings in terms of operating conditions and building characteristics. Even for the five buildings, error was no doubt introduced since metering was only conducted for the summer months and usage had to be allocated to the remaining non-summer months.

Statistical Analysis. Without fully appreciating the extent of random error contained in these priors for the 376 sites that had usable data, we tried a variety of specifications in an effort to make the most use of these sophisticated and expensive engineering data. The realization rates hovered around .35, which seemed implausibly small given what we thought at the time to be greatly improved/enhanced priors. Next, we allow only the 139 sites for which the priors were considered to be superior. Surprisingly, even here, we failed to obtain realization rates greater than .5, a number again considered implausible. Of course, none of these results is now surprising given the propagation of random error described above. Finally, more reasonable results were obtained when all engineering priors were replaced with dummy variables and a realization rate of .92 was achieved.

Conclusions

Based on these experiences, we have several conclusions regarding the future use of SAE models for those analysts who wish to avoid encounters of both the third *and* fourth kind. These recommendations are, for the time being, restricted to situations in which SAE models are used to estimate the impact of HVAC measures.

One should be very cautious in using engineering priors contained in utility DSM program tracking systems. Our experience and that of others suggest that engineering priors which reside in utility program tracking databases are riddled with error, a good deal of which is random error. Both statistical theory and Monte Carlo simulations have demonstrated that random measurement error in an independent variable produces biased estimates of β.

To eliminate engineering collection may be premature. Both Monte Carlo simulations and practical experience suggest that the challenges of SAE models,

[3] See Section V for a complete definition of Calibration, Cluster and Matched-Pair sites.

while not theoretically insurmountable, may be practically insurmountable owing to the high costs.

If one has used on-site data to detect and correct measurement error, one should be very careful in using the ratio of the improved priors to the original priors to adjust the original priors of those who have not received the more rigorous analysis. Attempting to leverage on-site using such a ratio may do very little to improve the original priors. This is the case since scaling the engineering prior, i.e., the independent variable, by this ratio, while changing the slope, β, by the inverse of the scaling factor, the magnitude of the random error and the resulting bias will remain the same.

One could estimate a model using *only* sites for which site-specific data had been collected and for which DOE2 models have been run. However, regression models can require two to four hundred sites given the size of the expected savings and/or regulatory requirements regarding sample sizes. This could be prohibitively expensive for most utilities.

The value of using dummy variables should not be underestimated. While not a very precise estimate of the expected savings, they are very reliable. That is, one could expect little measurement error since what we're measuring is the installation of the efficient equipment.

Finally, while quite ordinary, dummy variables are free.

References

Carmines, Edward G. and Richard A. Zeller, (1979), Reliability and Validity Assessment. Newbury Park: SAGE Publications.

Johnson, Jr., Aaron C., (1987), Marvin B. Johnson, and Rueben C. Buse. Econometrics: Basic and Applied. New York: Macmillan Publishing Company,.

Kennedy, Peter., (1992), A Guide to Econometrics. Cambridge, MA: The MIT Press.

Maddala, G. S., (1992), Introduction to Econometrics. Englewood Cliffs, N.J.: Prentice Hall.

Meyers, R.H. ,(1990), Classical and Modern Regression with Applications, Second Edition, Boston: PWS and Kent Publishing Company, Inc.

Sonnenblick, R. and J. Eto., (1995), A Framework for Improving the Cost-Effectiveness of DSM Program Evaluations. Berkeley, CA: Lawrence Berkeley Laboratory, University of California, LBL-37158: UC-1321.

Violette, Daniel, (1991) "Analyzing Data", Chapter 4 in: Handbook of Evaluation of Utility DSM Programs, (Eds. Eric Hirst & John Reed). Oak Ridge National Laboratory, Oak Ridge, Tennessee. ORNL/CON-336.

Violette, Daniel, Richard Brakken, Andy Schon, and Jerry Greer., (1993), “Statistically-Adjusted Engineering Estimates: What Can The Evaluation Analyst Do About The Engineering Side Of The Analysis?” Published in the *Proceedings of the 1995 Energy Program Evaluation Conference*.

Vine, Edward L. and Martin G. Kushler, (1995), “The Reliability of DSM Impact Estimates.” Published in the *Proceedings of the 1995 Energy Program Evaluation Conference*.