A *Monte Carlo* Based Comparison of Techniques for Measuring the Energy Impacts of Demand-Side Management Programs

Dean A. Schiffman, San Diego Gas and Electric Company

California's recently established DSM measurement protocols specifically prescribe a set of DSM measurement methods, including *Conditional Demand Analysis.* Recently, however, the *California Public Utilities Commission* approved a temporary (partial) exemption from the protocols for *San Diego Gas & Electric Company.* The exemption was based on evidence provided by SDG&E that it had developed a separate measurement method—termed *Identification Based Modeling (IBM)* or *Simplified Conditional Demand Analysis—which is (1)* at least equally effective in estimating the gross energy impact of DSM measures, (2) less susceptible to error, and (3) significantly less expensive to implement.

IBM begins with a conventional conditional demand regression model. However, to make the regression model less susceptible to misspecification and errors in data collection, the regression coefficients are strategically divided into two groups: those that will be held constant (when the model is eventually estimated using data), and those that will be allowed to vary across customers in the database. It is shown that allowing some regression parameters to vary across customers can dramatically reduce data collection requirements, with virtually no loss in the effectiveness of the regression model, resulting in significant reductions in DSM measurement expense. The conclusions are supported by a computerized *Monte Carlo* study that compared *IBM* and conventional conditional demand analysis under a variety of conditions.

Introduction

Many electric utilities currently face the task of applying regression techniques to customers' energy consumption data (used jointly in the regression with customer site data) to measure the impact of their commercial Demand-Side Management (DSM) programs. Since commercial DSM programs are often thought of at the end-use level (commercial *lighting* retrofit programs, for example, are quite prevalent, as are *space cooling* programs), these utilities often turn to regression techniques that likewise consider customers' energy consumption at the end-use level. (A prime example is the recently adopted *California DSM Measurement Protocols* which require utilities to derive estimates of lighting and space cooling consumption in addition to estimates of DSM savings for those same end-uses.) The most widely used regression technique of this sort is *Conditional Demand Analysis (CDA)* which can provide estimates of the end-use components of customers' energy consumption provided that, (1) adequate data are available and, (2) the mathematics of the regression equation is correctly specified.[1]

However, *as it is generally applied CDA* requires onerous amounts of customer-specific data and a high degree of mathematical specification. The data requirements not only imply substantial expense, but also significant customer aggravation with on-site visits for data collection. With respect to specification, it is generally true that regression equations which are heavily specified face a good chance of being mathematically *misspecified,* in which case the regression application either fails overall or, at best, provides biased estimates of the regression coefficients. This is born out in the case of *CDA* by the general view that while *CDA* has often been successfully applied to the residential sector, its performance in the commercial and industrial sectors (where energy consumption issues are more complex) is a question.[2]

The purpose of this paper is to demonstrate that while *CDA* provides a good starting point in any attempt at DSM measurement, more thought should be given to the separate phases of the DSM measurement task, so that certain elements of *CDA can* be exploited more effectively. As will be demonstrated, a more careful consideration of the separate phases of DSM measurement will point to ways in which the data requirements of *CDA can* be lessened, and to ways of reducing the chance of mathematical misspecification. Specifically, in the usual case where *CDA is* applied with the explicit goal of disaggregating customers' consumption by its end-use components, a large number of regression coefficients are automatically *jointly specified* and *estimated.* An alternative-presented here as *Identification Based Modeling,* or *IBM*— would be to clearly divide the mathematical specification of the *CDA* regression equation from the *econometric estimation* phase of the DSM measurement task. *(IBM* has been referred to previously as *Simplified Conditional Demand Analysis. See* Schiffman et al. 1993, and Schiffman and Engle 1993.) This division (depicted in Figure 1) creates an opportunity to first *identify* the theoretical role (within the DSM setting) of each specified regression coefficient before the task of estimation is begun; carefully doing so will go far in reducing the drawbacks of *CDA,* by eliminating the estimation of regression coefficients that are unneeded within the DSM measurement setting. In concrete terms, it will be shown that when the central task within the DSM setting is to estimate customers' DSM *savings* for (say) a *single* end-use such as lighting, it is unnecessary to estimate customers' *consumption* for *each* end-use (lighting, space cooling, space heating, ventilation, water heating, cooking, etc.), provided that role of each regression coefficient is carefully identified before the estimation phase is begun.

In summary, although *IBM* share *CDA* the same mathematical structure for the regression equation, they differ in their treatment of individual regression coefficients during the estimation phase of the DSM measurement task. As a result, it is important to compare the performance of these alternatives within the estimation phase. During 1993 San Diego Gas & Electric Company (SDG&E) performed a large-scale computer simulation study (a *Monte Carlo* study), based on an agreement with *California Public Utility Commission.* The study was designed to systematically compare the fundamental properties of *CDA* and *IBM* as DSM measurement tools. The results of this study demonstrate that under the best of conditions—where the regression equation is correctly specified and customer-specific data are available— *IBM* performs as well as *CDA* in estimating DSM savings, with far fewer data requirements. [3] Additional results show that when the regression equation is misspecified or when data are collected with errors, *IBM* significantly outperforms *CDA* as a tool for DSM measurement. These result will be presented here,

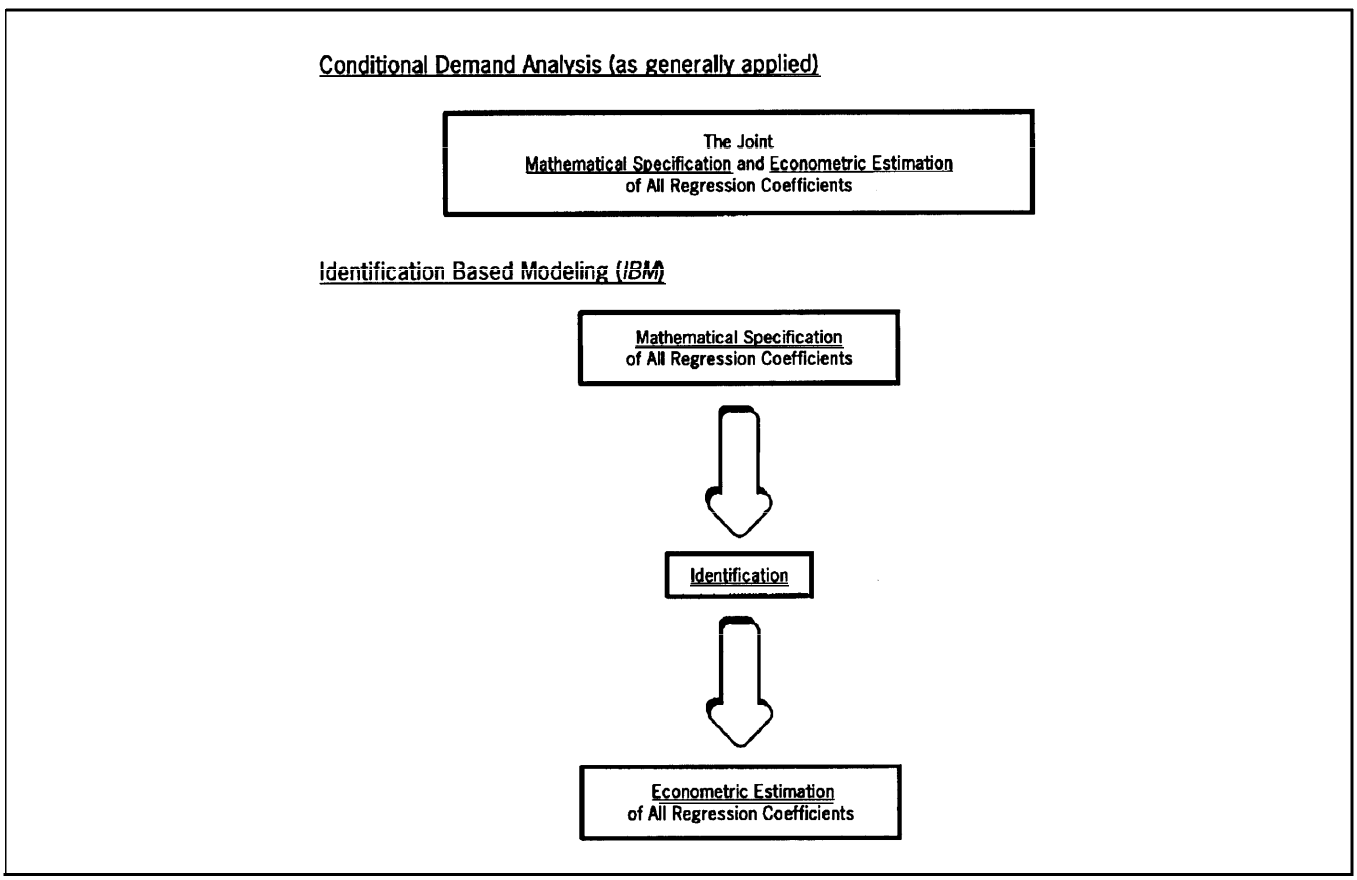

Figure 1. Conditional Demand Analysis Versus Identification Based Modeling

once the theoretical differences between *CDA* and *IBM* are established in forthcoming sections.

A Comparison of *Conditional Demand Analysis* and *Identification Based Modeling*

Specification of a Simple CDA Regression Equation (with No DSM)

Although it is far simpler than *CDA* regression equations that are generally specified in normal commercial applications, the following regression equation will serve well in both the theoretical discussion of *CDA* versus IBM and the portion of this paper containing the *Monte Carlo* results:

$$KWH_{jt} = \beta_{LIGHT}\left[(SQFT_j)\left(\frac{HRS_j}{1{,}000}\right)\right] + \beta_{COOK}\left[DCOOK_j\right] \quad (1)$$
$$+ \beta_{SPCOOL}\left[(BC_j \times SQFT_j)(CDD_{jt})\right] + \epsilon_{jt}$$

The regression equation has monthly electricity consumption for customer j during month t (KWH_{jt}) as the dependent variable. Given a simple three end-use setting with *interior lighting, cooking* (selected arbitrarily), and *space cooling,* there are three regressors (independent variables) on the right-hand side, each contained in square brackets. In turn, these regressors are based on:

$SQFT_j$ = building square footage for customer j,

HRS_j = monthly lighting hours-of-operation for customer j,

$DCOOK_j$ = a dummy variable which equals "1" when customer j has cooking ("0" otherwise),

BC_j = a factor representing the individual building characteristics for customer j, and,

CDD_{jt} = cooling degree-days for customer j during month t.

The usual random disturbance term ϵ_{it} has been included.

The regression coefficients are $_{LIGHT}$, $_{COOK}$, and $_{SPCOOL}$. The coefficient $_{LIGHT}$ is to be interpreted as lighting *watts per square foot.*[4] Cooking consumption $_{COOK}$ is simply a constant for those customers that have cooking. Space cooling is related to building characteristics BC_j (space cooling efficiency, wall/window/roof characteristics, etc.) and a weather variable CDD_{jt}; $_{SPCOOL}$ is a statistical adjustment factor.[5]

Specification of the CDA Regression Equation (with Lighting DSM)

If DSM lighting measures are installed within a portion of customers' facilities, the *CDA* regression equation becomes:

$$KWH_{it} = \text{right-hand side of Equation (1)}$$
$$+ \beta_{DSM}\left[(SQFTAFF_j)\left(\frac{HRS_j}{1{,}000}\right)(DDSM_{jt})\right] \quad (2)$$

where $SQFTAFF_j$ is the square footage that is affected by the lighting change ($SQFTAFF_j < SQFTj$), and $DDSM_{jt}$ is an econometric dummy variable which equals zero before the lighting change for customer j, and one thereafter. It follows that $_{DSM}$ is the *change in watts per square foot* of affected space.

Identification of Regression Parameters: *M-Identification* and *E-identification*

To this point, the term "identification" has been used without definition. Consider two separate terms: *M-identification* and *E-identification. M-identification will* be a loose concept based on whether the coefficient is simply important (iMportant) considering the goal of the regression analysis. *E-identification* will be based on the formal notion of econometric (Econometric) identification.

M-Idenlification (of a regression parameter): The regression parameter is important considering the goal of the regression analysis.

E-Identification (of a regression parameter): The independent variables of the regression are sufficiently structured so that the least-squares estimates of the regression coefficients are available.

It follows from the definitions that if a regression coefficient is *M-identified,* it must also be *E-identified* if least-squares estimates of the coefficient are to be obtained. However, if a parameter is *not M-identified,* it need not be *E-identified.* The usefulness of this will now be made plain using Equations (1) and (2).

Conventional CDA: Identification and Estimation

CDA was originally developed with the goal of disaggregating customers' energy consumption into its end-use components (see Parti and Parti 1980). If this *is* the goal of the regression analysis, by definition $_{LIGHT}$, $_{COOK}$, and $_{SPCOOL}$ of Equation (1) are *M-identified,* and must therefore be *E-identified.* This implies that there must be sufficient variation in all the regressors of Equation (1), so that the final data requirements for CDA are:

Data Requirements for Conventional *CDA* (no DSM):
$\{SQFT_j, HRS_j, DCOOK_j, BC_j, CDD_{jt}\}$

Within the DSM setting the data requirements are expanded, given that $_{DSM}$ of Equation (2) is now to be *M-identified:*

Data Requirements for Conventional *CDA* (with DSM):
$\{SQFT_j, HRS_j, DCOOK_j, BC_j, CDD_{jt}, SQFTAFF_j, DDSM_{jt}\}$

IBM: Identification and Estimation

If it is *not* the goal of the regression analysis to disaggregate customers' energy consumption into its end-use components, but rather the goal is to estimate DSM savings, it follows that *only* the coefficient $_{DSM}$ of Equation (2) is *M-identified.*

There are clear advantages in recognizing this fact. Each of the regression coefficients that are *not M-identified* (*namely* $_{LIGHT}$, $_{COOK}$, and $_{SPCOOL}$) are associated with *regressors that contain factors which do not depend on time;* these factors appear in square brackets in Equations (2)-(5). This establishes the following *customer-specific coefficients:*

$$B_{LIGHT,j} = \beta_{LIGHT}\left[(SQFT_j)\left(\frac{HRS_j}{1{,}000}\right)\right] \tag{3}$$

$$B_{COOK,j} = \beta_{COOK}\left[D_j^{cook}\right] \tag{4}$$

$$B_{SPCOOL,j} = \beta_{SPCOOL}\left[(BC_j \times SQFT_j)\right]. \tag{5}$$

It follows that Equation (2) can be rewritten as,

$$KWH_{jt} = B_j + B_{SPCOOL,j}\left[(CDD_{jt})\right] + \beta_{DSM}\left[(SQFTAFF_j)\left(\frac{HRS_j}{1{,}000}\right)(DDSM_{jt})\right] + \epsilon_{jt}, \tag{6}$$

where, $B_j = B_{LIGHT\,j} + B_{COOK\,j}$.

Equation (6) represents the final regression equation that is to be estimated under IBM. Note its fundamental properties. First, the data requirements have been reduced by $\{SQFT_j, DCOOK_j, BC_j\}$:

Data Requirements for *IBM* (with DSM):
$\{HRS_j, CDD_{jt}, SQFTAFF_j, DDSM_{jt}\}$

This reduction in data requirements stems directly from the identification issue. For example, $SQFT_j$ is *required* with the CDA setting, since it is only variation in which makes $_{LIGHT}$ *E-identified.* By contrast, in the DSM measurement setting $_{LIGHT}$ is not *M-identified,* which eliminates the need for variation in $SQFT_j$. (Note that it is the variation in $DDSM_{jt}$ that makes $_{DSM}$ *E-identified.)* In practice, the data elements $\{HRS_j, SQFTAFF_j, DDSM_{jt}\}$ would most likely be a part of the data-collection effort for any well established DSM commercial lighting program, while this would not be the case concerning the elements that have been eliminated, namely the *CDA* data requirements $\{SQFT_j, DCOOK_j, BC_j\}$.

Second, while in Equation (6), the coefficients $_{LIGHT}$, $_{COOK}$, and $_{SPCOOL}$ are not *E-identified,* neither are they *M-identified* within the DSM measurement setting; $_{DSM}$, of course, is both *M-identified* and *E-identified.* As a result, we see the fundamental difference between *CDA* and *IBM: IBM* will not yield estimates of the end-use components of consumption (as will *CDA,* if successfully applied) although it will yield estimates of DSM savings with fewer data requirements and a smaller degree of mathematical specification.

Of course, Equation (6) contains two sets of customer-specific coefficients: B_j and $B_{SPCOOL,j}$, j = 1,.. .J (in the case of J customers). To finally estimate the 2J + 1 regression coefficients, the exact form of the regression Equation would be (with a complete specification of the regressors):

$$KWH_{jt} = \sum_{j=1}^{J} B_j\left[D_{jt}\right] + \sum_{j=1}^{J} B_{SPCOOL,j}\left[(D_{jt})(CDD_{jt})\right] + \beta_{DSM}\left[(SQFTAFF_j)\left(\frac{HRS_j}{1{,}000}\right)(DDSM_{jt})\right] + \epsilon_{jt}, \tag{7}$$

where D_{jt} is a dummy variable which equals one when KWH_{it} is contributed to the regression by customer j, and zero otherwise. While constructing the 2J dummy variables of Equation (7) may require significant computer memory if J is large, modern hardware and software tools are generally sufficient. *SAS* PROC GLM, for example, is especially suited for constructing large numbers of dummy variables. The SAS code that would be written in order to estimate Equation (7) would resemble:

```
PROC GLM;
MODEL KWH = CUSTOMER   CUSTOMER*CDD
            SQFTAFF*INVIOOO*HRS*DDSM;
```

where the above variable names are based on the elements of Equation (7). The "CUSTOMER" and "CUSTOMER* CDD" portions of the code will yield estimates of the J customer-specific intercepts and the J customer-specific weather coefficients, respectively. The "INV1000* SQFTAFF*HRS*DDSM" portion of the code will generate an estimate of the single coefficient $_{DSM}$. Once the IBM framework is viewed in this way, another point of view becomes clear: Estimating Equation (7) amounts to estimating J customer-specific regression equations, subject to the restriction that the single regression coefficient for "INVIOOO*SQFTAFF*HRS*DDSM" is the same across customers. As a result, modeling restrictions have been minimized.

Testing the *CDA* Framework

In summary, IBM minimizes the constraints on regression coefficients that are present when *CDA* is applied. If within the setting of DSM measurement a choice must be made between *CDA*, with its heavy constraints on regression coefficients, and IBM, a less constrained framework, the natural course of action would be to formally test the constraints that are present within the *CDA* framework. Suppose, for example, that $_{SPCOOL}$ and $_{COOK}$ of Equation (1) may vary across customers, in which case there exists a set of coefficients $_{SPCOOL,J}$ and $_{COOK,J}$, j = 1,.. .J. Then one view is that *CDA* yields estimates under the joint hypotheses H_0-SPCOOL and H_0-COOK:

H_0-SPCOOL: $_{SPCOOL,j}$ = $_{SPCOOL}$, (for all j)

H_0-COOK: $_{COOK,j}$ = $_{COOK}$, (for all j where $DCOOK_j = 1$)

The hypotheses H_0-SPCOOL and H_0-COOK could be tested quite easily. The *Lagrange Multiplier (LM)* statistic (See Engle 1984) would function well as a test statistic in this case, since it depends only on estimates derived under the null hypothesis (i. e., the *CDA* regression estimates). Constructing the *LM* statistic would simply involve regressing the residuals from the *CDA* regression on a set of customer-based dummy variables (in addition to the original *CDA* regressors); based on this regression, the *LM* statistic could be calculated as: LM = (# observations) x R^2 If the *LM* statistic is "large" (compared to the appropriate Chi-square value) the null hypothesis would be rejected. (The simple intuition is that when the null hypothesis is false, the customer-based dummy variables which are not in the *CDA* regression equation end up in the disturbance term, and the least-squares residuals should reflect this.) If the hypothesis is rejected, the regression coefficients of H_0-SPCOOL and HO-COOK should be estimated at the customer level, as in *IBM.*

In general, the added flexibility of *IBM* becomes clear if we note that if either H_0-SPCOOL or H_0-COOK is false, only the *right-hand side* of the Equations (4) and (5) would be altered; the regression Equations (6) and (7) would maintain their present structure. For example, using *IBM*, the least-squares estimate of $B_{SPCOOL,J}$ would be an unbiased estimate of *either* $_{SPCOOL}[(BC_J \text{ x } SQFT_j)]$, when is true, or $_{SPCOOl,j}[(Bc_j \text{x} SQFT_j)]$, when H_0-SPCOOL is false.

The *Monte Carlo* Study

Description. In 1993, SDG&E applied to the California Public Utilities Commission for an exemption the California DSM measurement protocols. The protocols require estimates of the end-use components of customers' electricity consumption; the company's application was based on its success in applying *IBM* to its commercial lighting retrofit program.

To support its application, SDG&E offered to undertake a *Monte Carlo* study to compare the performance of conventional *CDA* and *IBM*. A *Monte Carlo* study is a computer-based study that simulates the application of these alternative techniques to actual data. In general, *Monte Carlo* methods have been used extensively in Economics, Business, and Statistics for many years. In 1984, the *Handbook of Econometrics* surveyed the applications in economics (see Hendry 1984), although the list of such studies is much longer today. These methods are now becoming so computationally inexpensive that they are often advocated as a companion to econometric model building. The company's proposal to undertake a *Monte Carlo* study was finally sanctioned by the CPUC in May of 1993 *(CPUC* decision D.93-05-063). The main results are given here. Details can be found in Schiffman and Engle 1993.

The computer simulated "world" of the *Monte Carlo* study provides several advantages. First, this world can be duplicated at will, so that issues such as which DSM measurement technique should be adopted can be analyzed under a variety of conditions; the "real world" takes place only once. Second, in the real world, the *true* numerical value for that item which is being estimated (e.g., energy savings from a DSM program) is unknown; as a result, there is no *known* standard by which alternative analytical tools can be judged. In a *Monte Carlo* study, a known value for the item which is being estimated can be built directly into the computerized framework, and alternative analytical tools can be judged in terms of their ability to detect this known value.

The *Monte Carlo* study was based on specifying the numerical values for each element of Equation (2), for a hypothetical group of 120 (J = 120) customers that would form the basis for separate *CDA* and *IBM* analyses. Although Equation (2) is relatively simple, it is complex enough to test whether adding the constraints that are present in the *CDA* framework (together with the data that are required to support these constraints) adds accuracy to the regression analysis, relative to *IBM.*

The overall study was based on simulating *one thousand* pairs of analyses (each pair being one *CDA* analysis and one *IBM* analysis); each pair is said to constitute one "iteration" within the study. With each iteration, each customer is randomly assigned new values for {$SQFT_j$, HRS_j, DCOOKj, CDD_{jt}, SQFTAFFj, $DDSM_{jt}$} (BC_j was suppressed). The weather variable CDD_{jt} (which naturally varied from month to month, across 36 months) was also changed with each iteration. Most important, each iteration contained new values (across customers and time) for the disturbance term ϵ_{it} of Equation (1). It is the randomness of the disturbance term which obfuscates the underlying structure of the regression equation, so that the alternative DSM measurement methods *(CDA* and *IBM)* can be compared in terms of their ability to overcome this element of the model. The overall numerical specification is described in Table 1.

Based on Equations (1) and (2), and the specifications in Table 1, a typical customer who has cooking should be expected to consume approximately 33,000 kWh per month (after the lighting change):

$$KWH_{jt} = 33{,}525 = (2.76)\left[(25{,}000)\left(\frac{500}{1{,}000}\right)\right] + 2{,}000 + (.000022)\left[(25{,}000)\left(\frac{9{,}000 + 22{,}000}{2}\right)\right] - (1.84)\left[\left(\frac{10\% + 90\%}{2}\right)(25{,}000)\left(\frac{500}{1{,}000}\right)\right] \tag{8}$$

Table 1. Numerical Specification of the *Monte Carlo* Study

Item Type	Item	Description of Numerical Specification
Regression Coefficients	LIGHT	2.76
	DSM	-1.84
	COOK	2,000 kWh per month
	SPCOOL	.000022
Customer Characteristics	$SQFT_j$	Assigned to customer within the range: 25,000 ± 6,250 square feet
	HRS_j	Assigned to customer within the range: 500 ± 40 hours per month
	$SQFTAFF_j$	Assigned to customer within the range: 10% - 90% of $SQFT_j$
	$DCOOK_j$	60% of customers were assigned cooking as an end-use.
	$DDSM_{jt}$	Consumption data were simulated over thirty-six months. Lighting changes took place within months 13 - 24.
Weather	CDD_{jt}	This variable was based on *ASHRAE* weather standards (see Schiffman and Engle 1993, and *ASHRAE* 1989 for details) for southern California, and ranged from approximately 9,000 to 22,000 with summertime peaks.
Random Disturbance	ϵ_{jt}	Assigned to customers using a random number generator (uniform distribution) so that the mean of ϵ_{jt} was zero, and its standard deviation ranged (across customers and iterations) from 10 - 30% of the average value for expected monthly consumption. The variance was assumed to be constant for any one customer within an iteration.
Monthly Consumption	KWH_{jt}	Constructed over 36 months, based on the above numerical specifications.

The example in Equation (8)—which serves only as a check on the reasonableness of the model—points to a breakout of monthly consumption of roughly three-fourths for lighting (before the lighting change), one-fifth for space cooling, and about 5% for cooking. In light of these specifications, the study can generally be thought of as an analysis of medium-sized office buildings. Note that the two lighting coefficients imply a two-thirds reduction in watts per square foot (1. 84/2.76 = 2/3); these numbers are consistent with an aggressive commercial lighting retrofit program that focuses on delamping and relamping.

Results. A series of 1,000 computer simulations were run based on the numerical specifications of Table 1. In particular, weighted least-squares estimates of the parameter were obtained for both the *CDA* specification of Equation (2) and the *IBM* specification of Equation (6); the weights of the regressions were based on estimates of the error-variances,

$$E\left(\epsilon_{jt}^2\right) = \sigma_j^2$$

Figure 2 gives the results of the 1,000 iterations in the case where the regression equation is correctly specified (that is, when the regression equation that is specified in the estimation phase is, in fact, Equation (2)). Not surprisingly, both estimators are obviously unbiased, since each distribution has a mean estimate for $_{DSM}$ of very close to -1.84 watts. More important, it is clear that the two estimators are extremely close in terms of accuracy. The two distributions are virtually congruent, although the tails of the *IBM* distribution are slightly more pronounced; the standard deviation of the *CDA* distribution (0.011) is 80% of the standard deviation of the *lBM* S-CDA distribution (0.014). This is an important result in light of the fact that the *IBM* estimator does not depend on the data elements {$SQFT_j$, $DCOOK_j$, BC_j}. In other words, the added accuracy that would, in practice, be associated with these data elements is minimal. Moreover, this small degree of added accuracy is predicated on the accuracy of these data elements (e.g., Can the square footage of a building be consistently and accurately measured?), and the exact mathematical specification of the of the *CDA* equation. We can now consider the properties of the two estimators under conditions of data-collection errors and modeling misspecification.

It is well known that when regression models contain regressors for which data are collected with errors, the resulting *errors-in-variables bias* will impact the estimation process, and that parameter estimates will tend to be biased toward zero. Figure 3 shows the results for the *CDA* and *IBM* models in a simulated case of *errors-in-variables bias* (based on an additional 1,000 iterations of the model). The results are based on the case where,

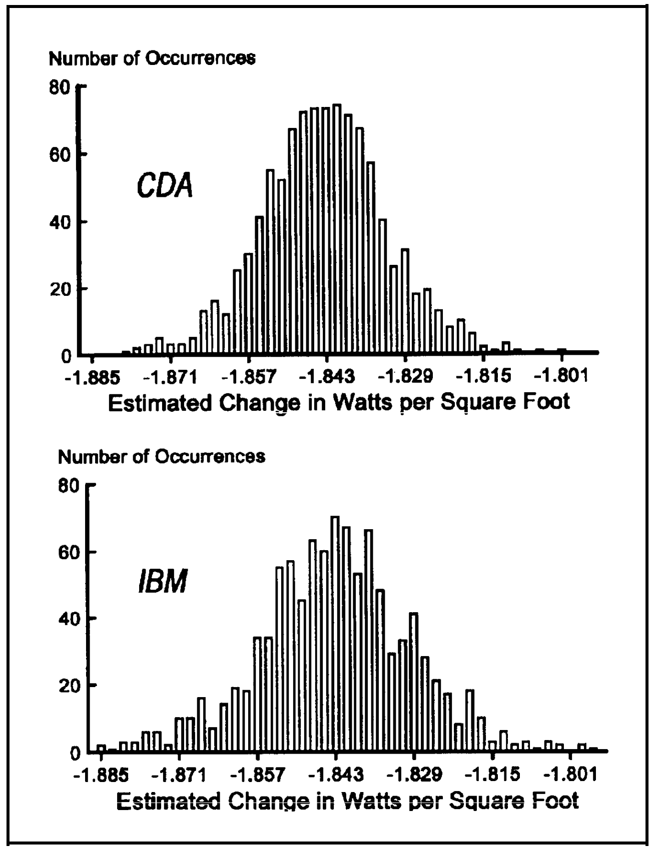

Figure 2. Monte Carlo Results: The Distribution of Estimation Results from 1,000 Iterations, when the Regression Equation is Correctly Specified During Estimation

during the data-collection and estimation phases, the regressor $SQFT_j$ is overestimated or underestimated by 0-20% (this implies that the building square footage on average will be recorded correctly, but in any one instance will be recorded with error by at most 20%). Figure 3 shows the significant *errors-in-variables bias* that is associated with the *CDA* estimator in this case, as well as a dramatic decrease in accuracy. (The resulting estimates are clearly biased toward zero away from the true value for the measure savings parameter $_{DSM}$, -1.84 watts.) Most important, it is clear in Figure 3 that since Equation (6) does not depend on $SQFT_j$, the *IBM* estimator is unaffected by the data collection errors. An alternative view is that the weighted least-squares estimate of the customer-specific space cooling coefficient in Equations (5) and (6) will certainly constitute an efficient, unbiased estimate of $_{SPCOOL}[(BC_J \text{ X } SQFT_j)]$ regardless of the data errors, since the *regressor* itself does not depend on square foot-age. In short, the customer-specific coefficient *implicitly* "collects" data on square footage data. In general, these results support the notion that the cost of estimating the individual end-use elements that are contained in Equation (2) will most likely be a substantial *errors-in-variables bias* and a significant decrease in

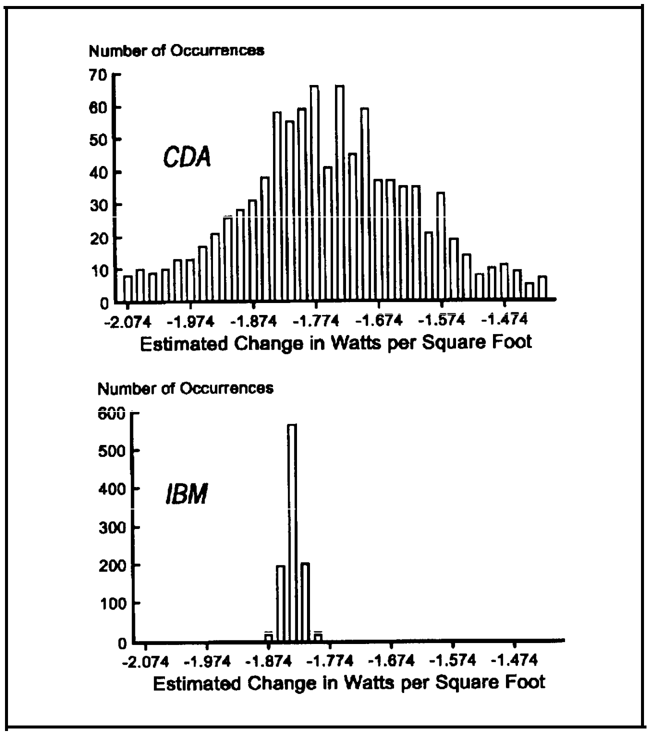

Figure 3. Monte Carlo Results: The Distribution of Estimation Results from 1,000 Iterations, when Square Footage is Over/Under Estimated by 0-20% During Estimation

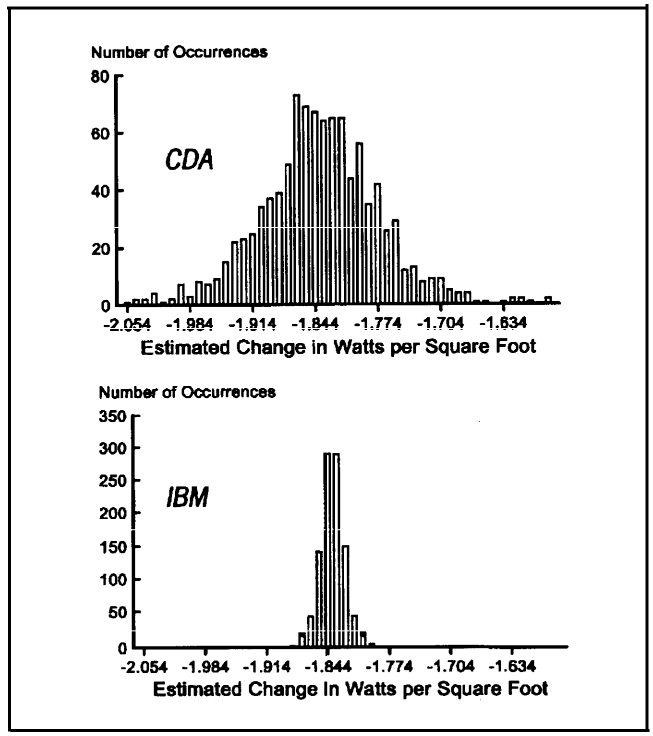

Figure 4. Monte Carlo Results: The Distribution of Estimation Results from 1,000 Iterations, when a Minor End-Use is Omitted from the Regression Equation During Estimation

accuracy. Note, also, that if a *CDA* regression equation more complicated than Equation (2) were specified, there would be an even greater chance for errors in data collection.

Problems can also occur when the regression equation is mathematically misspecified. Figure 4 shows the results (based on an additional 1,000 iterations) of omitting the cooking end-use during the estimation phase of *CDA*. The variance of the C-CDA estimator is clearly dramatically increased, although the associated bias is not conspicuous (undoubtedly due to the small consumption that is associated with this end-use). It is also clear that since the *IBM* Equation (6) does not depend on the cooking indicator $DCOOK_j$, the accuracy of the *IBM* estimator is unaffected by the misspecification. Once again, an alternative view is that the weighted least-squares estimate of the customer-specific intercept B_j in Equation (6) will certainly constitute an efficient, unbiased estimate of $B_{LIGHT,j}$ + ${}_{COOK}[DCOOK_j]$, regardless of the misspecification, since the *regressor* itself does not depend on $DCOOK_j$. In short, the customer-specific intercept *implicitly* "collects" data on cooking. These results show that the accuracy of *CDA is* significantly impacted by *misspecification error.* Note again that if a *CDA* regression equation more complicated than Equation (2) were specified, there would be an even greater chance for misspecification.

Windfall Estimates from *IBM*

The intuitive notion associated with *IBM* is that it is not necessary to estimate the *level* of energy consumption by end-use when the fundamental task is to estimate the *change* in consumption for, say, a single end-use. Why, for example, should the analyst estimate the level of cooking consumption in commercial office buildings if the basic task is to estimate the change associated with lighting retrofit work? Given *IBM* it is clear that changes is energy consumption for a particular end-use can generally be estimated without estimating the level of consumption for that end-use or others.

However, SDG&E has found some unexpected benefits in applying *IBM* which pertain to estimating the level of lighting consumption. To understand these benefits, imagine a simple case where a commercial lighting retrofit program is based on a single lighting measure, that of delamping (removing) 4-foot fluorescent lamps from lighting fixtures; the typical case could be one where one lamp is removed from every 4-lamp fixture. In this case, based on the results presented here, it is clear that *IBM* could produce estimates of the energy savings from delamping, and that these estimates could easily be constructed on a per-lamp basis. Yet this delamp estimate could just as well serve as an estimate of the energy consumption for the *remaining* lamps (for the three remaining lamps per fixture, given the typical case just described). In other words, in the case where for a particular end-use a DSM measure involves removal (as in delamping), the level of consumption can be estimated given information concerning the relationship between the level of consumption and the DSM measure. As an example of this, in applying *IBM* to its commercial lighting retrofit program SDG&E first carefully defined a detailed set of DSM lighting measures (based primarily on lamp/ballast combinations) and, as in the above example, a subset of these measures were associated with delamping. SDG&E has shown how the delamping estimates can be used to estimate the level of lighting consumption for commercial sites (see Schiffman et al. 1994).

Summary and Conclusions

Conditional Demand Analysis (CDA) is certainly an appropriate technique when the task is to disaggregate commercial customers' energy consumption into its end-use components. However, when the task is to estimate the impact of DSM measures for (say) a particular end-use, there are advantages in using *Identification Based Modeling (IBM). IBM* (which begins with the mathematical specification for the regression equation that is found in the *CDA* framework) allows for a careful consideration of the issue of parameter *identification* before the final form of the regression equation (the regression equation that is to be *estimated) is* established. The value of *identification* lies in significantly reduced data requirements, and a lower susceptibility to data-collection errors and errors in mathematical specification.

Acknowledgments

The author would like to thank Professor Robert F. Engle for his many helpful comments concerning this work, as well as the econometricians associated with the major California utilities, the California Energy Commission, and the California Public Utilities Commission.

Endnotes

1. For an overview of *CDA*, see Parti and Parti 1980, Lawrence and Parti 1984, Sebold and Parris 1989.

2. Referring to *CDA* as a "pure econometric approach," Regional Economic Research 1991, has stated "Because of the complexity of commercial building systems and the diverse nature of occupant behavior, the pure econometric approach does not work for commercial sector applications. "

3. The results that will be presented here are actually a subset of a larger set of results (see Schiffman and Engle 1993) that have led the *California Public Utilities Commission* to approve a temporary (partial) exemption from the California measurement protocols for SDG&E *(CPUC* decision D. 93-10-063, May 19, 1993).

4. In reality, *watts per square foot* will not be constant across customers. A more realistic application of *IBM* would be to clearly define a set of lighting "measures" whose watt-savings could be viewed as constant. See Schiffman, et al. 1993, for such an application.

5. The regression equation could be viewed as containing a lighting/space-cooling interaction if lighting hours were contained within space cooling hours-of-operation and space cooling was of sufficient duration. The original study also contained terms for heat gain from building occupants and office equipment (see Schiffman and Engle 1993).

References

American Society of Heating, Refrigerating and Air Conditioning Engineers. 1989. *ASHRAE Handbook Fundamental Z-P Edition,* Atlanta, GA.

Engle, Robert F. 1984. "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics." *Handbook of Econometrics.* Griliches and Intrilligator, North-Holland.

Hendry, David F. 1984. "Monte Carlo Experimentation in Econometrics." *Handbook of Econometrics.* Griliches and Intrilligator, North-Holland.

Lawrence, A.G. and M. Parti. 1984. *Survey of Conditional Energy Demand Models for Estimating Residential Unit Energy Consumption Coefficients.* Report EA-3410, EPRI, Palo Alto, CA.

Parti, Michael and C. Parti. 1980. "The Total and Appliance Specific Conditional Demand for Electricity in the Household Sector" *The Bell Journal of Economics,* 2(1).

Regional Economic Research. 1991. *Commercial Sector Regional Data—1986 NBECS Market Profiles.* Energy Information Administration.

Schiffman, Dean A., Athena Besa, Andrew Sickels, and J.C. Martin. 1993. *Commercial/Industrial Energy Efficiency Incentives: Lighting Retrofit Estimation of Gross Energy-Demand Impacts.* MIAP-92-P50-S01-R320, San Diego Gas & Electric, San Diego, CA.

Schiffman, Dean A., and Robert F. Engle. 1993. *Appendix Z Simulation Study, Comparison of Alternative Methods for Measuring the Gross and Net Energy Impacts of Demand-Side Management Programs (with Addendum).* MIAP-93-P99-S01-R330, San Diego Gas & Electric, San Diego, CA.

Schiffman, Dean A., Patrick Kirkland, and Andrew Sickels, 1994. *Commercial/Industrial Energy Efficiency Incentives: Combining Fixture-Metering Data and Customer-Billing Data for Gross Impact Measurement.* MIAP-93-P50-903-416, San Diego Gas & Electric, San Diego, CA.

Sebold, F.D. and K.M. Parris. 1989. *Residential End-Use Energy Consumption: A Survey of Conditional Demand Estimates.* Report CU-9487, Research Project 2547-1, EPRI, Palo Alto, CA.