# Practical Issues in Using Customer Billing Data for Evaluating Commercial, Industrial, and Large Multifamily Conservation Programs

Scott Pigg, Wisconsin Energy Conservation Corporation

Customer billing records have been a valuable resource for evaluating demand side management programs. The data are readily available for large groups of customers, and can be obtained without intruding upon the customers or facilities of interest. Many evaluations of residential energy efficiency programs have successfully used customer billing data to estimate aggregate program effectiveness.

But commercial, industrial, and large multifamily energy efficiency programs pose challenges to program evaluators who would like to make use of customer billing data. A single program may embrace facilities that vary in size by several orders of magnitude, have multiple accounts per facility, and comprise a variety of businesses and energy consumption patterns. These characteristics create obstacles to program evaluators, ranging from the mundane (but crucial) problem of identifying the proper accounts to analyze, to the complex task of adjusting diverse consumption patterns for variations in the weather, to the difficult issues of identifying valid comparison groups.

This paper identifies and addresses several important practical obstacles in using billing data to analyze these programs. The paper is intended to promote awareness and discussion among evaluators of the challenges faced in analyzing billing data for these sectors.

## Introduction

In contrast to residential dwellings, commercial and industrial facilities range widely in size, consume gas and electricity for a wide variety of uses, and do so in patterns that are not always easily discerned. Multifamily dwellings, though residential in nature, also come in widely different sizes and utility metering configurations.

The strong suit of customer billing data is its ubiquity: it is the only source of measured consumption data that is available for every customer in a utility's service territory. Customer billing records can also be obtained retrospectively; this is an important advantage for many rebate programs in which participants are not known until after they have already installed conservation measures. Fels and Reynolds [1991] argue cogently that the ubiquity of billing data offers opportunities for standardization of whole-facility savings analysis that embraces the interacted effects of all energy efficiency efforts in facilities.

The weak points of customer billing data are its coarseness and comprehensiveness, however. Most billing data represent monthly meter reads, which precludes assessing load shape impacts. Moreover, customer billing data represent energy use for all end-uses that are attached to the billing meter, which usually includes consumption for many end uses that are not of interest for the evaluation of a particular program.

This creates two challenges in using these data: (1) discerning the effect of the program among the considerable "noise" from a multitude of end-uses; and (2) ensuring that any average change in consumption is the result of the program, and not the result of an aggregate change in consumption for another reason or among the other end-uses that are present in the billing data.

In one way or another, evaluations that use billing analysis boil down to trying to statistically discern an average change in energy usage between periods before and after participation. To eliminate changes in consumption because of nonprogram effects, two groups are studied: a group of program participants, and a group of similar customers who did not participate. Models that attempt to explain some of the usage variation in the billing data can get quite complicated (e.g., Parti and Rogers 1991), as can strategies for dealing with systematic differences between participants and non-participants (e.g., Train and Ignelzi 1987).

This paper is not about the theoretical aspects of specific models or techniques for conducting billing analysis-based evaluation. Rather, it addresses some practical barriers and issues that largely transcend the specific methods or models used to assess program savings. These issues involve characteristics of the customers, their billing data, and the energy efficiency programs in which they participate, all of which present challenges to the evaluator in estimating program energy savings, regardless of the method used.

## Issues

I have chosen to address five practical issues that often arise in attempting to measure energy savings from energy efficiency programs that address facilities in these sectors. These are:

- identifying the right billing data, and defining the units of sampling and analysis;
- dealing with the large variation in energy usage and program impacts among facilities;
- sample size, attrition, and sample representativeness issues;
- selecting a comparison group; and,
- weather-normalizing consumption.

Several of these issues have been discussed by Schuler [1990] in reference to assessing savings in institutional buildings. My perspective is more focused on utility incentive programs that involve a larger and more varied population of customers.

In detailing these issues, I will occasionally refer to data from an evaluation of a recent conservation competition in Madison Gas & Electric (MG&E) Company's service territory (Pigg et al. [1991]). The MG&E Competition involved gas and electric conservation measures installed by MG&E and three competitors in large (over 5-unit) multifamily buildings as well as commercial and industrial facilities. As such, it provides a good overview of these sectors and the problems encountered in measuring program impacts from customer billing data.

### Issue 1: Identifying the Right Billing Data and Defining the Units of Sampling and Analysis

The simple but unpleasant truth (for evaluators) is that utility billing systems are designed to bill customers efficiently, not collect data for program evaluation. Working within a system designed for another purpose can impede our ability to both identify the data we wish to study and analyze it in the manner we prefer.

Multifamily buildings and commercial facilities are metered in a multitude of ways. In some cases there are individual meters for separate businesses or apartments; in other cases, businesses within a single building share a common meter. Sometimes multiple buildings are represented on a single billing meter. All-electric apartment buildings are often entirely individually metered, with separate meters for common areas such as hallways. Apartment buildings with gas service may be master-metered, individually metered, or have a combination of master metering for space heating or domestic hot water and individual metering for apartment ranges.

After-the-fact evaluation of program impacts can be considerably impeded by having to assess which meters are relevant to the evaluation, and which represent consumption for end-uses that are not of interest. About half of the 1,840 service addresses involved in the MG&E Competition had multiple accounts, and overall, 13,642 accounts were associated with these service points. Most of the facilities with multiple accounts had to be scrutinized by hand in order to identify the accounts of interest, a task that required several hundred person hours of time. Nadel and Ticknor [1989] report similar difficulties in linking program databases with customer billing systems, and Okumo [1990] documents the massive data preparation requirements that were needed to evaluate energy efficiency measures undertaken in all-electric multifamily buildings.

Recording customer account numbers as part of routine program tracking can overcome some of these barriers, but keep in mind that "account" is a financial term, not a metering term. An account may embrace multiple billing meters and even multiple facilities. At some utilities, a portion of the account number represents the physical premise or service point: this portion remains constant even if the customer changes, and is the preferred identifier for program tracking. At other utilities, the account number itself provides no information about the location of the facility and is liable to be completely changed if a business or apartment building changes hands. In addition, reledgering or changes to meter reading routes occasionally render old account numbers invalid. These changes may be difficult to trace by computer.

Customer name and service address are often recorded for participants in conservation programs, but these have their own problems in linking to a mainframe billing system. Service address can be searched to find all accounts or service points at a particular address, but will often yield more accounts and meters than are needed. And some buildings have more than one service address. Addresses must be listed in a standardized manner that is compatible with the billing system, or no match will be made. Searching on customer name can be a quagmire for multifamily and commercial customers, because accounts are often listed under management companies that may hold hundreds of accounts at separate locations.

A more fundamental issue, however, is that utility billing systems impose structural constraints on how we sample and analyze data. End-uses, conditioned spaces, and even whole buildings are already bundled in configurations that we must live with. This can blur the definition of the basic unit of analysis, particularly if we target only the billing meters that are affected by the program. In master-metered buildings, the usage being studied may represent a multitude of end-uses and affected spaces; or, by virtue of individual metering, consumption may represent a single end-use. Keating and Blachman [1987], Wolfe and McAllister [1989] and others report dropping facilities from study because of a single billing meter representing multiple buildings. While this has the potential to bias the results of an evaluation, retaining facilities with different metering arrangements means there may be substantial differences in what a "building," "firm" or "facility" actually represents. If the number of configurations are not large, it may be possible to define broad categories of metering arrangements, and analyze each separately.

In addition, if consumption for multiple meters or accounts is aggregated to represent a facility, problems can arise when we attempt to match non-participant facilities to the participants using the customer billing system. Schemes for matching participants with non-participants (which I discuss in more detail later) are usually run at the account level. This makes it difficult to find a comparison match for a facility whose consumption is defined as the sum of multiple accounts.

Analysis of multifamily program impacts often presents a choice of whether to use buildings or apartments within a building as the unit of analysis. The choice may be determined by the dominant metering configuration of participants (master-metered or individually metered), and may be influenced by the type of measures installed in the program.

The problems outlined above may or may not be a significant barrier to billing analysis; it depends on the nature of the program that is being studied and the nature of the utility billing system. For the three multifamily, commercial, and industrial programs I have worked with, the majority of facilities were cleanly distinguished by a single meter, but the minority that had more complicated metering arrangements required an inordinate amount of time to categorize, and created problems in matching non-participants to participants.

## Issue 2: Dealing With a Large Variation in Usage and Impacts Among Facilities

There are several related characteristics of multifamily, commercial, and industrial customers and energy efficiency programs that influence the ability to statistically discern program effects from other sources of variation in the billing data. First, there tends to be a tremendous variation in the size of customers who participate in commercial and industrial energy efficiency programs (and to a lesser extent, in multifamily programs). Gas and electric usage for the programs I have examined have been log-normally distributed, as Figure 1 demonstrates for the MG&E Competition. This means that while the logarithm of usage follows a familiar bell-shaped distribution, usage itself is skewed towards some very large customers. Note how usage varies over three to four orders of magnitude between participants. Usage for residential customers rarely varies over more than one or two orders of magnitude.

Second, while both large and small customers tend to show similar *percentage* changes in usage from year-to-year, the *absolute* variability in usage is much more for large customers. This can be seen by examining gas usage for facilities in the MG&E Competition for the two years immediately preceding participation in the program. Figure 2 shows the percentage change in gas usage between 1987 and 1988 versus the usage level in the first year (usage is annualized, but not weather normalized here).[1] In percentage terms, large and small customers show a similar year-to-year fluctuation in usage. (There is a general increase in usage, because the weather was colder in the second year.)

But this similar percentage variability translates into a much larger absolute variation for large customers (Figure 3). A fluctuation of 10% represents only 100 therms for a 1,000-therm customer, but represents 10,000 therms for a 100,000-therm customer. If we assume that the percentage variability in usage is roughly constant with increasing usage, then the variability of usage in absolute terms increases linearly with usage level (the increase looks exponential in Figure 3 because of the log scale for usage). Since the change in usage is how we normally estimate savings, this result means that the random error associated with assessing the average

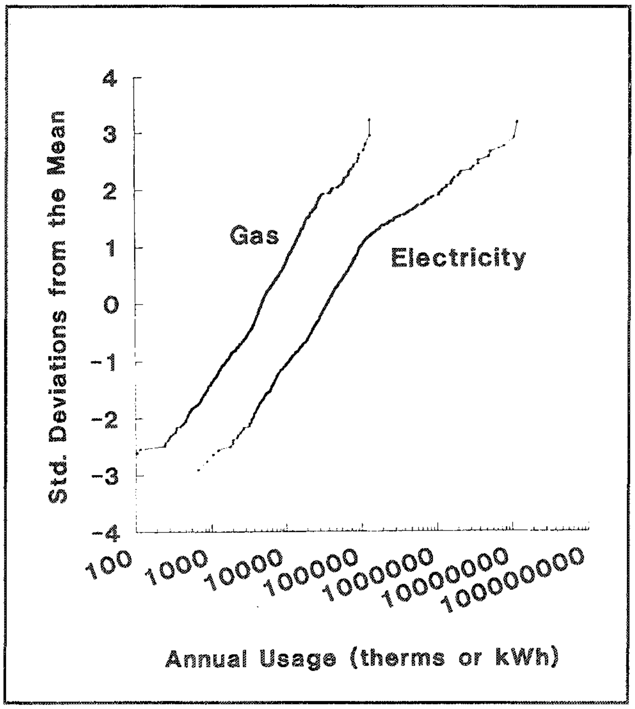

*Figure 1. Log-Normal Probability Plot of Gas (n=1044) and Electric (n=910) Usage for Facilities in the MG&E Competition. A log-normal distribution will plot as a straight line. Note how consumption varies over 3-4 orders of magnitude.*

savings increases with usage. This effect can be seen in Table 1, which shows the mean change in usage, the standard error, and the coefficient of variation by usage quartile for the MG&E gas customers.

Third, it is not unusual that a substantial portion of the total impact of programs in these sectors comes from a few large projects undertaken by large customers. Figure 4 shows that when facilities are ranked from largest to smallest in terms of the estimated energy savings, the top 10% of projects are expected to provide 50% (for gas) to 80% (for electricity) of the total estimated savings for the program for the MG&E Competition.

The more the impacts of a program are concentrated among a few large customers, the less amenable the program will be to statistical billing analysis. The situation is exacerbated by a tendency for large customers to be idiosyncratic in terms of the nature of their business and the technologies employed under the program. In some cases, it may be necessary to separate the analysis of a few large projects from that of the majority of smaller projects, and use a case study approach for the former.

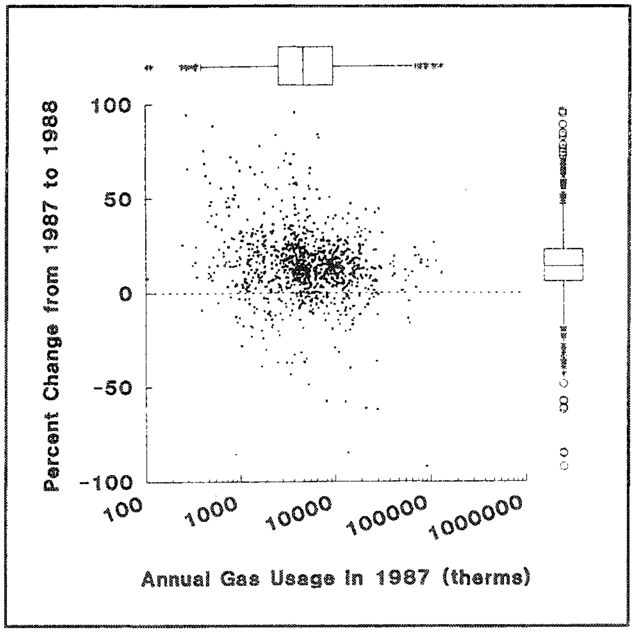

*Figure 2. Percentage Change in Annual Gas Usage for 1044 Facilities in the MG&E Competition for the Two Years Prior to Participation Versus Usage in the First Year (Log Scale). Large customers are similar to small customers in their usage variability.*

Energy savings for very large projects deserve extra attention anyway, and can be estimated using techniques such as enhanced engineering estimates based on site visits or end-use metering. Reiwer and Spanner [1991] used this approach in evaluating the performance of a small number of large industrial projects. Statistical billing analysis should be reserved for larger samples of customers that are more homogeneous with respect to consumption and savings levels.

When the situation is less extreme, the participant population can be stratified so that a larger proportion of the total sample is allocated to higher usage customers. The key requirement for this strategy is that the variable being used for stratification be known for all participants, not just the study sample. This argues for routinely tracking some measure of facility usage in program databases. Even when stratified by size, however, the precision of the overall billing analysis may be limited by the number of participants available for study in the larger strata.

In general, it is vital that impact evaluation in these sectors begins with an assessment of how usage and impacts are distributed among the participant population.

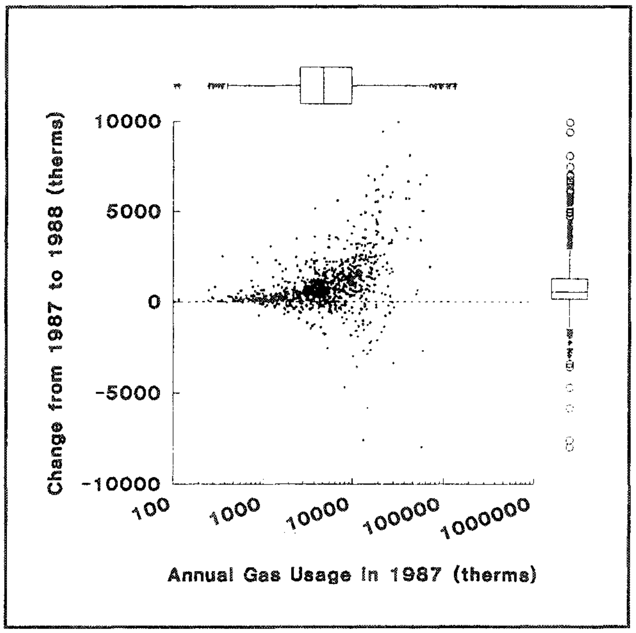

*Figure 3. Therm Change in Annual Gas Consumption for 1044 Facilities in the MG&E Competition Between the Two Years Prior to Participation Versus Usage in the First Year (Log Scale). By virtue of their size, large customers have substantially more variation in consumption.*

## Issue 3: Sample Size, Attrition, and Sample Representativeness Issues

***The Tradeoff Between Sample Size and Facility-Level Data.*** The statistical ability to resolve an average change in consumption depends on how many facilities we can study and the unexplained variability in consumption from year to year. Sometimes, by virtue of a large participant population and automated access to billing data, we can analyze large groups of facilities. If all data transfer and analysis is performed electronically, the marginal cost of analyzing each additional facility is small. It is therefore feasible and advisable to simply analyze all available program participants, rather than draw a sample from the participant population.

In other circumstances, sample sizes may be small, but we may be able to gather additional information on each facility that helps explain some of the variation in energy consumption and thereby increase the precision in the estimate of program induced savings. Rarely can we have it both ways, however. And having a program tracking system with detailed facility-level data does not resolve this trade-off if nothing is known about the comparison group, or if there is no information about changes that occurred in the participant sample *after* participation.

This creates a choice between having a large but noisy sample or a smaller sample with auxiliary information that may help reduce the unexplained variance in consumption. It is usually easier to make an *a priori* assessment of the statistical precision of the former approach than the latter. Knowing the size of the participant population and the distribution of usage, we can make an estimate of the year-to-year variability in usage, and assess the statistical precision for different sample designs for both participants and non-participants.

It is harder to know beforehand the usefulness of auxiliary information in helping explain variance in the billing data. In using modelling techniques that attempt to explain the variability in usage, Parti and Rogers [1991] report success for evaluating a commercial and industrial program with 376 customers (and end-use metering of 60 customers), while Nadel and Ticknor [1989] describe difficulties in applying these models.

Table 1. Change in Annual Gas Usage Between 1987 and 1988 for MG&E Competition Participants by 1987 Usage Quartile

| Quartile | n | Minimum Usage (1987) | Maximum Usage (1987) | Mean Change (1987-1988) | Standard Error | Coefficient of Variation |
|---|---|---|---|---|---|---|
| 1 | 261 | 42 | 2540 | 327 | 35 | 10.5% |
| 2 | 261 | 2552 | 4687 | 630 | 43 | 6.8% |
| 3 | 261 | 4702 | 9592 | 878 | 69 | 7.8% |
| 4 | 261 | 9629 | 132066 | 1975 | 441 | 22.3% |
| Total | 1044 | 42 | 132066 | 952 | 114 | 12.0% |

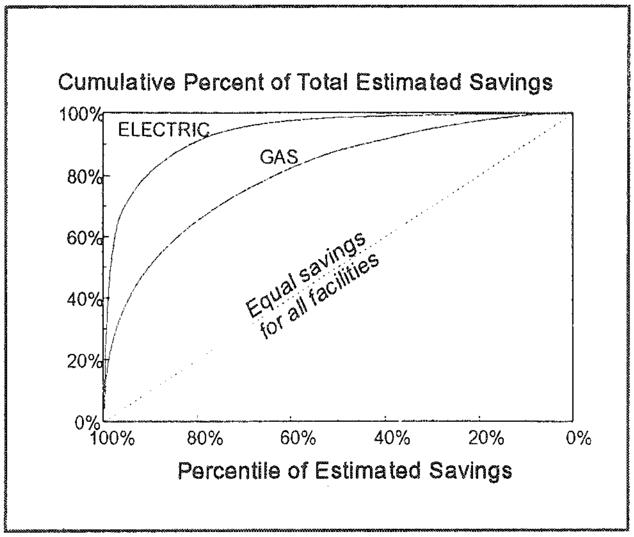

*Figure 4. Cumulative Estimated First-Year Energy Savings as a Function of Facility Ranking in Terms of Estimated Savings. The top 10% of facilities account for 50% (gas) to over 80% (electric) of all estimated impacts.*

If the potential sample size is large and access to billing data is automated, a reasonable dual approach may be to perform a simple pre/post analysis on a large sample of participants and non-participants, then follow up this analysis with a more detailed and complex analysis of a smaller subsample. This has the advantage of allowing better assessment of the representativeness of the smaller sample with respect to the larger one. It can also be used to target specific subgroups of customers for whom the initial analysis suggests the need for closer scrutiny.

*Finite Population Correction.* When a large proportion of the participant population is included in the analysis, evaluators sometimes employ the finite population correction factor (fpc) on the grounds that the results are more certain than the usual statistics (which assume an infinite population) indicate.[2] If the fpc is to be used, several points need to be borne in mind:

- With the fpc, statistical precision estimates reflect only the uncertainty in measuring the immediate population under study; if the target population is in fact a larger group of similar customers who may participate in the near future, the confidence in generalizing the evaluation results will be overly optimistic.

- In general the fpc can only be applied to the participant group; non-participants are usually sampled from what is still an essentially infinite population. This reduces the effect of the impact of the fpc on estimates of net impacts, since the unadjusted uncertainty of the comparison group dominates the overall uncertainty.

- The fpc assumes that measurements of savings are made without error; in fact, there is uncertainty in the estimates of normalized annual pre- and post-participation usage, and hence in savings. This uncertainty is usually small in relation to the uncertainty that arises from differences in savings between facilities, but disregarding this measurement error can lead to a false apparent statistical precision when the fpc is used.

- Finally, almost all billing analysis studies (indeed, all observational studies) have to some extent or another biases that are not reflected in the statistical precision estimates. These biases may be minor in relation to an unadjusted estimate of precision, but may invalidate precision estimates that are estimated using the fpc.

Unless one can still justify using the fpc after working through this list, it is better not to apply it.

*Sample Attrition.* Sample attrition is always a concern in billing analysis studies, and occurs in two varieties: attrition that is forced on us, and attrition that we choose to enforce. The former type of attrition arises from the timing of program participation (we can only study facilities that have an adequate history of pre- and post-participation billing data), and missing or incomplete billing histories. Attrition from these sources can be severe, but cannot be avoided. One can only analyze the representativeness of the facilities that are left, and perhaps re-weight the results to better reflect the population of interest.

On the other hand, enforced attrition is what happens when facilities are excluded from analysis, not because the data are lacking, but because they are anomalous, are not expected to contribute useful information, or considered unreliable in some way. Examples of this type of attrition include:

- excluding facilities that do not show a good fit to weather normalization models (Miller and Vedadi [1989]);

- excluding master-metered facilities whose billing consumption includes multiple buildings (Keating and Blachman [1987], Wolfe and McAllister [1989])

- excluding facilities for which the expected impact of the conservation measure is very small in relation to total billing consumption (Rogers [1989], Pigg et al. [1991]); and

- dropping facilities that are remodelled, have a change in schedule, or a change in tenancy (Wolfe and McAllister [1989]).

These screens may improve the statistical precision of savings estimates, but they also have the potential to bias the results of the analysis against the type of facilities that are dropped. In particular, as I discuss later, I do not recommend dropping facilities on the basis of poor correlation with the weather. The second and third examples above represent attempts to exclude facilities that are not expected to contribute useful information to the analysis because the measures implemented are dwarfed by other consumption sources.

Tenancy changes and remodelling create a particularly thorny issue for billing analysis and sample attrition. In non-residential settings, a tenancy change can substantially change the amount and pattern of energy use. Business expansion and remodelling can have a similar effect. The available evidence indicates that facility changes that affect energy consumption are fairly frequent (Hickman and Steele [1991], Petersen [1990, 1991], and may often coincide with program participation.

Including these businesses can add a lot of highly variable usage data to a billing analysis. But excluding them reduces the study sample to a subgroup of the participant population. The choice of how to handle these changes may depend on the size of the study group. If the available sample size is large, the higher level of usage variability due to facility changes may be surmountable. If only a small sample of participants and non-participants is available, there may be no choice but to analyze a smaller group of stable facilities, and recognize that the ability to generalize the results is reduced.

But including or excluding facilities on the basis of stability presumes knowledge of tenancy and remodelling changes. This may not be the case if we have a large data set of billing information with little supporting information about each facility. In this case, we are faced with a collection of outlier facilities that show large changes in energy consumption. One can blindly trim a data set of outliers, but this tends to overstate the true precision of the savings estimate, since statistically we just pretend that the outliers don't exist.

In general, I have found the distribution of the year-to-year percentage change in billing usage to be leptokurtic (that is, excessively "peaky" compared to a Normal distribution). This effect can be seen in the long tails of the box plot at the right of Figure 2. This distribution shape suggests that across any two time periods we normally deal with a mixture of two kinds facilities; (1) stable facilities for which the year-to-year change in usage arises from variation in the operation of a relatively static set of energy-using equipment, and (2) facilities that undergo major changes in appliance holdings or a structural change of some kind, which show a much larger variability in usage.

*Assessing Attrition Bias.* Because attrition of one form or another is always present in billing analysis studies (and indeed in all evaluation methods that lack true random assignment to treatment and control groups), it is important to analyze how the final study groups differ from the participant and nonparticipant populations that we are studying. Cochran [1983] states that "the investigator may do well to adopt the attitude that, in general, estimates of the effect of a treatment or program from observational studies are likely to be biased." The question is whether these biases are large enough to worry about.

One potential stumbling block to assessing bias is that we can only measure it for data that exists for both the sample and the population. For the kinds of data that are typically collected for program participants, this means assessing the geographic, market segment, energy usage, and conservation measure representativeness of the study group against the participant population. If substantial biases are revealed, the study groups can be post stratified on these variables to better reflect the populations.

This raises the question of what constitutes a "substantial" bias. Blasnik[3] recently pointed out that significance tests are often inappropriately used to assess bias. One typically calculates the mean value of some variable for the sample and the population, and concludes that the sample is not significantly different from the population unless the probability from a t-test is less than, say 5%. But used in this way, the more strict one makes the rejection level of a t-test, the worse a bias has to be in order for it to show up as significant. Some biases may be of concern when there is a 25% probability that the measured difference between sample and population could occur by chance. Other biases may not be important even if significant at less than a 1% level. Patton [1982] notes that a classic evaluation pitfall is "assuming that statistically significant results are always practically significant, and that statistically insignificant results are always practically insignificant." The assessment of bias must go beyond a mechanical significance test, and be based on the evaluators judgement, for which statistical tests provide only part of the information.

## Issue 4: Comparison Group Selection Issues

Most evaluations that use billing analysis include a group of non-participants. Because of the diversity of businesses among commercial and industrial customers, it can be difficult to identify a non-participant group that is a good proxy for the participant group.

Practically speaking, it is usually difficult to be sure how much of the net effect from a billing analysis is really due to the program even when a comparison group is employed. The fact that participants chose to be in the program while non-participants did not can imply differences in energy use that may masquerade as an effect of the program. This well-known problem of self-selection bias is a concern for all observational studies that lack truly random assignment to treatment and control groups, and has been discussed in detail elsewhere (e.g., Violette and Ozog [1989]).

Some evaluations compare the participant study group to usage changes for an entire rate class in a utility's service territory. This avoids the issue of finding a sample of nonparticipants, but introduces the potential for more severe self-selection bias. The issue I will focus on here is on the practical difficulties of identifying a reasonably well-matched group of non-participants under a matching scheme designed to find a sample of non-participants that are comparable to participants.

A comparison group needs to be matched to the participant group in at least two important ways: the size and type of facilities represented, and energy usage levels (matching on energy use intensity, such as kWh/ft$^2$ is preferable, but this data is not usually available for the entire pool of nonparticipants). However, there are some barriers to implementing this type of matching procedure. The first goes back to the issue of how we define what we are analyzing. A scheme that uses the customer billing system to match participants and non-participants must operate using the units that the customer billing system is based on--typically customer accounts or service points. If an evaluation is based on a definition of a facility that can embrace multiple customer accounts or service addresses, it can be impossible to match comparison facilities using the customer billing system. In the MG&E Competition evaluation, we were forced to match using customer accounts, even though about 20% of the participant facilities embraced multiple accounts. In these cases, the result was a single participant facility being matched to multiple non-participant facilities.

Standard Industrial Classification (SIC) codes are often used to match business types. But not all utilities include an SIC code in their billing systems, and when they are available, SIC codes may be missing or incorrectly coded for many customers. And even using reliable SIC codes does not guarantee a good match in all cases between each participating and non-participating customer. Even at the four-digit level[4], two businesses with the same SIC code may be very different in how they operate and use energy.

Despite these difficulties, my feeling is that using an imperfect market segment identifier such as SIC code is better than using none at all. In general, matching based on SIC will pair a church with a church and a restaurant with a restaurant. If close matching is a concern and the participant sample is small, multiple non-participants can be matched to each participant from the customer, and more rigorous (and labor intensive) screening applied to find a non-participant that best represents each participant.

More troublesome than finding the best match among smaller commercial customers, is finding any match at all among the few larger participants. These facilities may simply be unique to a particular service territory. This again argues for separate handling of these projects, apart from the main billing analysis of a larger sample of smaller customers.

Matching on customer energy usage can be performed in a couple of ways. If the participant group is stratified by usage, the non-participant sample can be drawn to provide a target sample size in each usage stratum. Alternatively, each participant can be matched with one or more non-participants that are similar in consumption level. It is important, though, that usage matching be based on consumption that occurs *before* program participation. Many utility billing systems maintain computer averages of usage per day or per heating degree day, which are usually updated every six months to a year. Unless the matching occurs soon after participation, these estimates will not be suitable for matching participants and non-participants, except as a very rough guide.

## Weather-Normalization Issues

Many evaluations are conducted in two-stages. In the first stage, monthly billing data are corrected for weather variation and annualized, giving normalized annual consumption (NAC) estimates of pre- and post-participation consumption. The second stage of analysis involves estimating program impacts from these NAC estimates.

There are two levels to the weather normalization question: first *whether*, and second, *how*? In some cases, weather normalization may not be necessary. Energy consumption for space conditioning may be a small fraction of total usage; as weather dependent consumption decreases as a proportion of total usage, so does the need for weather-normalization. In this case one may simply annualize monthly billing data, and rely on the comparison group to account for aggregate changes in consumption due to weather variation, as Coates [1991] did in an evaluation of a commercial conservation program in Seattle. In practice, the relative dominance of space conditioning loads is likely to vary from facility to facility within a single program, so weather normalization may be important for some facilities but not for others.

How to normalize non-residential billing data for weather variation in an automated way is a difficult problem. Depending on the nature of the business, the climate, and whether we're looking at gas or electric consumption, usage may show a space heating load, a space cooling load, both loads, or neither. Space conditioning may constitute a large percentage of total energy use or may be small in relation to other end-uses. And consumption may be seasonally variable in a way that looks like weather dependence but really isn't.

Reynolds et al. [1990] use a case study of electricity usage in a New Jersey shopping mall to discuss the use of a "sieve" procedure for classifying small commercial buildings according to whether their consumption patterns indicate the presence or absence of space heating and cooling. Using a similar procedure, I classified gas and electric accounts in the MG&E Competition according to various energy signatures.[5] The results for electric accounts (almost all gas accounts showed a heating signature) are compared in Figure 5. The classifications are similar, despite the fact that the New Jersey dataset comprised facilities in a single shopping mall, while the MG&E dataset is made up of a variety of commercial and industrial facilities. The main difference is that more heating-only facilities show up in the MG&E study group.

Only a minor percentage of accounts show both space cooling and heating loads, which is fortunate, since simultaneous heating-and-cooling normalization is more difficult, and software for this purpose is not generally available. Of course, these results may vary regionally.

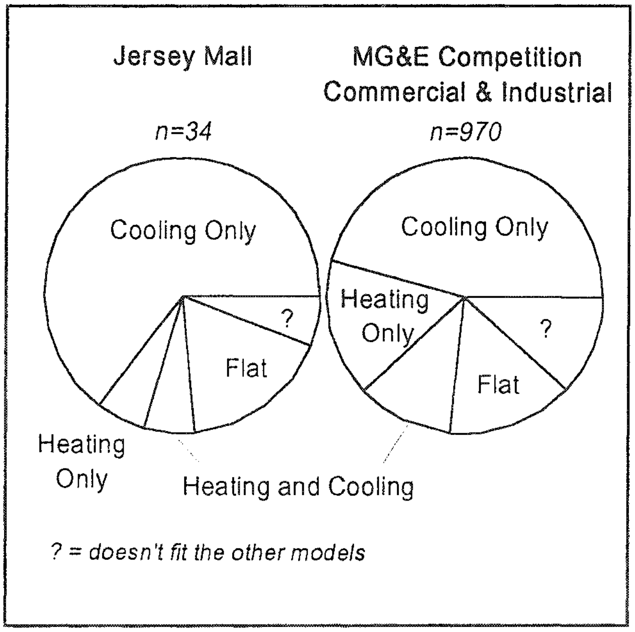

*Figure 5. Classification of Electric Accounts by Weather Dependence Category for Businesses in a New Jersey Shopping Mall (Reynolds et al. [1990]), and for Commercial and Industrial Facilities in the MG&E Competition*

One particularly troubling point is that in performing this procedure for both pre- and post-participation periods, 48% of 2,987 electric accounts were classified in one category in one period and a different category in another period. Fels and Reynolds [1991] report similar findings for commercial buildings in the Northeast. The question becomes whether one should force the same model on both periods of consumption, or whether separate models should be used. I used whichever model was least restrictive. For example, if an account showed flat consumption in one period, but cooling dependence in another period, I used the cooling model for both periods, since a cooling model can accommodate flat consumption. More investigation is needed on the stability of weather dependence in commercial buildings, though.

My feeling is that when a two-stage approach is used for billing analysis, some protocol should be used to weather-normalize consumption for buildings whose consumption clearly shows a strong correlation with outdoor temperature. But we should not be overly concerned with capturing and weather-normalizing all weather-dependent consumption. So long as the same procedure is followed for both participants and non-participants, changes in consumption due to uncorrected weather sensitive consumption should show up in both the treatment and comparison groups and be taken out of estimates of net energy savings. In the same vein, I do not recommend dropping facilities that show a poor fit to weather normalization models. Energy usage in these sectors cannot be expected to be dominated by space conditioning loads to the extent that consumption will always show a weather dependence.

## Conclusions

In order for billing analysis-based evaluation to be successful in these sectors, program tracking systems need to have clearly defined linkages into customer billing systems, and contain facility type and energy usage information, as well as information about the estimated impacts of the energy efficiency measures that were implemented.

Samples need to be designed to accommodate the large variation in facility energy consumption and program impacts, and some large customized projects may not be amenable to statistical billing analysis.

To some extent there is a trade-off between having a large but noisy sample, and having a smaller sample with additional supporting information. Choosing an approach to take may largely be determined by the size of the pool of available participants and non-participants that can be studied.

Non-participants need to be carefully matched with participants, and all study groups should be assessed for representativeness with respect to the target population.

And finally, a protocol for identifying and handling weather dependent usage may be needed depending on the facilities, conservation measures and fuel being studied. But weather normalization should be viewed as a tool for assessing those facilities for which space conditioning loads are clearly dominant; it should not be prerequisite for inclusion in the analysis.

Energy efficiency programs in these sectors range from a few highly customized projects in unique large facilities to high-volume rebate programs in a targeted population of customers. Analysis of customer billing data cannot resolve all evaluation questions in all programs, but if applied under the right circumstances, it can provide valuable and informative insights about program impacts in multifamily, commercial, and industrial facilities.

## Endnotes

1. The 1987 year was actually defined to comprise consumption from October 1, 1986 to September 30, 1987. The 1988 year was defined similarly. Accounts with at least 270 days of usage data are shown. A few outliers are not visible in the figures.

2. The finite population correction factor, as applied to the standard error of an estimate is given by $(1\text{-}n/N)^{1/2}$, where n is the sample size, and N is the size of the population (Cochran [1977]).

3. Michael Blasnik, GRASP, Philadelphia, Pennsylvania, personal communication

4. SIC codes are like postal zip codes; the first digit makes broad distinctions such as between manufacturing, retail and agricultural activities, and each successive digit subdivides these classifications into finer subcategories.

5. Reynolds et al. used the Princeton Scorekeeping Method (PRISM) software for weather normalization (Fels [1986]). We used our own software, which uses the same model and a similar fitting algorithm as the heating/cooling-only versions of PRISM. We modeled simultaneous heating and cooling using a piecewise linear, dual change-point model of consumption versus average outdoor temperature using a non-linear statistics program.

## References

Coates, B. 1991. "Energy Savings and Cost-Effectiveness in the Commercial Incentives Pilot Program," Proceedings of the 1991 Energy Program Evaluation Conference, pp. 368-373, Chicago, IL.

Cochran, W. G. 1983. *Planning and Analysis of Observational Studies*, John Wiley & Sons, New York.

Cochran, W. G. 1977. *Sampling Techniques*, 3rd edition, John Wiley & Sons, New York

Fels, M. 1986. "PRISM: An Introduction." *Energy and Buildings*, Vol. 9, Nos. 1 & 2, pp. 5-18.

Fels, M. F., and C. L. Reynolds. 1991. "Toward Standardizing the Measurement of Whole-Building Energy Savings in DSM Programs," Proceedings of the 1991 Energy Program Evaluation Conference, pp. 75-85, Chicago, Illinois.

Hickman, C. and T. Steele. 1991. "Building Site Visits as Supplement to Program Evaluation," Proceedings of the 1991 Energy Program Evaluation Conference, pp. 174-180, Chicago, IL.

Keating, K. M., and S. Blachman. 1987. "In Search of an Impact: An Evaluation of an Institutional buildings Program," Proceedings of the 1987 Energy Program Evaluation Conference, Vol. 1, pp. 107-116, Chicago, IL.

Miller, J. R. and A. Vedadi. 1989. "Evaluation of the Utah Institutional Conservation Program: Preliminary Results", Proceedings of the 1989 Energy Program Evaluation Conference, pp. 297-303, Chicago, IL.

Nadel, S., and M. Ticknor. 1989. "Electricity Savings from a Small C&I Lighting Retrofit Program: Approaches and Results" proceedings of the 1989 Energy Program Evaluation Conference, pp. 107-112, Chicago, IL.

Okumo, D. L. 1990. "Multifamily Retrofit Electricity Savings: the Seattle City Light Experience," Proceedings from the ACEEE 1990 Summer Study on Energy Efficiency in Buildings, Vol. 6, pp. 6.119-6.130, American Council for an Energy Efficient Economy, Washington, D.C.

Parti, M., and E. Rogers. 1991. "Conditional Demand Analysis of Commercial and Industrial Customers to Determine the Effects of a Demand-side Program," Proceedings of the 1991 Energy Program Evaluation Conference, pp. 323-327, Chicago, IL.

Patton, M. Q. 1982. *Practical Evaluation* Sage Publications, Beverly Hills, CA.

Petersen, F. J. 1990. "Remodel and Tenancy Changes: Threats to the reliability of Commercial Conservation Savings," Proceedings from the ACEEE 1990 Summer Study on Energy Efficiency in Buildings, Vol. 3, pp. 3.165-3.172, American Council for an Energy Efficient Economy, Washington, D.C.

Petersen, F. 1991. "Changes to IBP Buildings" Proceedings of the 1991 Energy Program Evaluation Conference, pp. 41-48, Chicago, Illinois.

Pigg, S., J. Schlegel, and J. Ford, "Impact Evaluation of a Multisector Conservation Competition," Proceedings of the 1991 Energy Program Evaluation Conference, pp. 328-336, Chicago, IL.

Reynolds, C., P. Komor, and M. Fels. 1990. "Using Monthly Billing Data to Find Energy Efficiency Opportunities in Small Commercial Buildings," Proceedings from the ACEEE 1990 Summer Study on Energy Efficiency in Buildings, Vol. 10, pp. 10.221-10.232, American Council for an Energy Efficient Economy, Washington, D.C.

Riewe, S., and G. E. Spanner. 1991. "Performing Impact Evaluations in Industrial Retrofit: the Energy $avings Plan Program," Proceedings of the 1991 Energy Program Evaluation Conference, pp. 485-491, Chicago, IL.

Schuler, V. 1990. "Measuring the Impacts of Energy Efficiency Measures in Institutional Buildings with Billing Data: a Review of Methodological Issues," Proceedings from the ACEEE 1990 Summer Study on Energy Efficiency in Buildings, Vol. 6, pp. 6.155-6.165, American Council for an Energy Efficient Economy, Washington, D.C.

Train, K., and P. Ignelzi. 1987. "Evaluation of a Conservation Program for Commercial and Industrial Firms," *Energy*, Vol. 12, No. 7, 1987.

Violette, D., and M. Ozog. 1989. "Correction for Self-Selection Bias: Theory and Application," Proceedings of the 1989 Energy Program Evaluation Conference, pp. 241-250, Chicago, IL.

Wolfe, P., and L. McAllister. 1989. "The Industrial Lighting Incentive Program: Process and Impact Evaluation," Proceedings of the 1989 Energy Program Evaluation Conference, pp. 99-106, Chicago, IL.