# Appendix B

# Nonsampling and Sampling Errors

## Appendix B

# Nonsampling and Sampling Errors

## Introduction

All the statistics published in this report are estimates of population values, such as the total floorspace in U. S. commercial buildings. These estimates are based on reports from representatives of a randomly chosen subset of the entire population of commercial buildings. As a result, the estimates always differ from the true population values.

The differences between the estimated values and the actual population values are of two types, nonsampling errors and sampling errors. Nonsampling errors are differences that would be expected to occur in all possible samples, or in the average of all estimates from all possible samples.

The first four sections (Nonresponse, Annual Consumption and Expenditures, Annual Peak Electricity Demand, and Additional Data Notes) that follow this introduction describe some of the sources of nonsampling error and how the survey is designed and conducted to minimize such errors. Nonsampling errors can result from: (1) inaccuracy in the data collection due to questionnaire design errors, interviewer error, respondent misunderstanding, and data processing errors; (2) nonresponse for an entire sampled building (unit nonresponse); (3) nonresponse on a particular question (item nonresponse); and (4) incomplete coverage due to deficiencies in the sampling frame. The section "Nonresponse" provides an overview of the procedures used to handle unit nonresponse and the item nonresponse associated with the building characteristics portion of the survey. (For a detailed discussion of these procedures, see Appendix C, "Nonsampling and Sampling Errors," in *Commercial Buildings Characteristics 1992*; April 1994, DOE/EIA-0246(92)). The consumption and expenditures featured in this report were based on monthly billing records submitted by the buildings' energy suppliers. The section, "Annual Consumption and Expenditures" provide a detailed explanation of the procedures used to obtain annual consumption and expenditure estimates from the bills, as well as the procedures used to handle partial or completely missing data. The peak electricity demand estimates in this report were also based on the monthly billing data, as described in the section, "Annual Peak Electricity Demand." The fourth section dealing with nonsampling error is titled, "Additional Data Notes," and discusses special problems encountered when reconciling building and supplier reports on energy sources used, gas transported for the account of others, demand-side management programs, and primary energy consumption for electricity.

The last section, "Estimation of Sampling Errors," describes how the sampling error is estimated and presented for statistics given in this report. Sampling errors are random differences between the survey estimate and the population value that occur because of the particular sample that was selected by chance. The sampling error, averaged over all possible samples, would be zero, but since there is only one sample for the 1992 CBECS, the sampling error is nonzero and unknown for the particular sample chosen. The sample design permits sampling errors to be estimated.

Most unit nonresponse cases were caused by a respondent's unavailability or refusal to participate in the survey. Item nonresponse resulted when the respondent did not know, or, less frequently, refused to give the answer to a particular question. Unlike the sampling error, the nonsampling error's magnitude cannot be estimated from the sample data. For this reason, avoiding biases at the outset is a primary objective at all stages of survey design and field procedures. The wording and format of the survey questionnaires, and the quality control built into the data collection, receipt, and processing operations were all designed to minimize these sources of error. For a discussion of the questionnaire design, interviewer training and data control, see Appendix A, "How the Survey Was Conducted."

# Nonresponse

## Unit Nonresponse

The response rate for the 1992 Building Characteristics Survey portion of CBECS, as reported in Appendix A, was 91.1 percent. That is, of the 7,282 buildings eligible for interview, 8.9 percent did not respond at all to the Building Characteristics Survey. This rate was similar to that for the 1989 CBECS and represents an extremely low unit nonresponse rate for a voluntary survey of this length and complexity.

Weight adjustment was the method used to reduce unit nonresponse bias in the survey statistics. The CBECS sample was designed so that survey responses could be used to estimate characteristics of the entire stock of commercial buildings in the United States. Weight adjustment resulted in basic sampling weights (base weights) that related the sampled buildings to the entire stock of commercial buildings. In statistical terms, a base weight is the reciprocal of the probability of selecting a building into the sample. A base weight can be understood as the number of actual buildings represented by a sampled building: a sampled building that has a base weight of 1,000 represents itself and 999 similar (but unsampled) buildings in the total stock of buildings.

To reduce the bias from unit nonresponse in the survey statistics, the base weights of respondent buildings were adjusted upward, so that the respondent buildings would represent not only unsampled buildings but also nonrespondent buildings. The base weights of respondent buildings were multiplied by the adjustment factor "A," defined as the sum of the base weights over all buildings selected for the sample divided by the corresponding sum over all respondent buildings. Respondent weights remained nonzero after weight adjustment. Nonrespondent weights were set to zero, because they were accounted for by the upward adjustment of respondent weights.

Unit nonrespondents tended to fall into certain categories. For example, nonresponse tended to be higher in the Northeast than in the Midwest. To reduce nonresponse bias as much as possible, adjustment factors were computed independently within 119 subgroups according to characteristics known from the sampling stage for both responding and nonresponding buildings. These characteristics included the general building activity, the approximate size of the building, Census region, and metropolitan versus nonmetropolitan area.

## Item Nonresponse—Building Characteristics

Item nonresponse is the type of nonresponse that occurs when an item (or several items) is missing in an otherwise completed questionnaire. Nonresponse in the Building Characteristics Survey was imputed to allow publication of *Commercial Buildings Characteristics 1992*, the companion volume to this report. The Energy Suppliers Survey consisted of four distinct data collections (electricity, natural gas, fuel oil, and district heating/cooling surveys) to obtain 1992 consumption information for buildings in the Building Characteristics Survey. Partial and complete nonresponse in the Energy Suppliers Survey are discussed in this section under "Annual Consumption and Expenditures."

The companion volume contains item nonresponse rates for many of the building characteristics used to present estimates in this report. Nonresponses to items in the Building Questionnaire were treated by a technique known as "hot-deck" imputation. In hot-decking, when a certain response is missing for a given building, another building, called a "donor," is randomly chosen to furnish its reported value for that missing item. That value is then assigned to the building with item nonresponse (the nonrespondent or "receiver").

To serve as a donor, a building had to be similar to the nonrespondent in characteristics correlated with the missing item. This procedure was used to reduce the bias caused by different nonresponse rates for a particular item among different types of buildings. Characteristics that were used to define "similar" depended on the nature of the item to be imputed. The most frequently used characteristics were: principal building activity, floorspace category, year constructed category, and Census region. Other characteristics (such as type of heating fuel, type of heating and cooling equipment and the responses to particular items in the 1986 CBECS for those buildings that were surveyed in 1986) were used for specific items.

As in the 1986 and 1989 surveys, the 1992 CBECS used a vector hot-deck procedure. With this procedure, the building that donated a particular item to a receiver also donated certain related items if any of these were missing. Thus, a vector of values, rather than a single value, is copied from the donor to the receiver. This procedure helps to keep the hot-decked values internally consistent, avoiding the generation of implausible combinations of building characteristics.

### Special Imputations for 1992 CBECS

In 1992, due to natural disasters, there were large areas that were inaccessible to interviewers and consequently, no interviews could occur at buildings in those areas. Because these buildings were clustered in a few areas, they were not adequately represented by buildings elsewhere. Therefore, it was decided that the unit nonresponse adjustment procedure would not be the optimal way to compensate for these buildings. Instead, in those few areas, all of the building characteristics, including eligibility, were imputed based on information available from the 1992 sample listing stage and from the 1986 survey. To hot-deck values for a particular item, all buildings were first grouped according to the values of the matching characteristics specified for that item. Within each group defined by the matching variables, donor buildings were assigned randomly to receiver buildings.

## Annual Consumption and Expenditures

This report presents estimates of energy consumption and expenditures in commercial buildings during calendar year 1992. These estimates were computed from the annual consumption and expenditures determined for each building in the CBECS sample. However, these "annual" values were not obtained directly from the suppliers for the buildings. Rather, energy suppliers provided monthly billing data, which were used to calculate calendar year consumption and expenditures for each building, according to the procedures described in this section. Also described in this section are the imputation procedures used in cases where the energy supplier survey data were unavailable or inadequate.

To assure that calendar year 1992 consumption would be completely accounted for, the data requested from suppliers were bills covering the period from December 1991 through January 1993. These bills formed the basis for the annual energy consumption and expenditures estimates published in this report.

### Billing Data: Ideal and Reality

The basic consumption and expenditures data were reported for each building by billing period. Ideally, the data for each continuous-delivery energy source (electricity, natural gas, and district heating and cooling) used in each sampled building should have been in the form of complete records for consecutive billing periods either totally or partially contained in calendar year 1992, covering exactly the energy consumed within the sampled building. The data for the discrete-delivery energy source (fuel oil) should have been in the form of complete data records for all deliveries during 1992. For both continuous- and discrete-delivery energy sources, the delivered energy source should have been used entirely within the sampled building.

In practice, though, the billing data often covered more or less square footage than just the sampled building's square footage, or did not match the target time frame, calendar year 1992. There were several common types of discrepancy between the bill coverage and the ideal of a single building and fixed time frame.

- Bill coverage included days in 1991 and 1993 as well as calendar year 1992. This was the typical situation for a complete billing record. Very rarely would one billing period begin on January 1 and another end on December 31, 1992.

- Bill coverage spanned at least a 1-year period, but did not include all of 1992. In most such cases, the time frame covered by the bills extended from the middle of 1992 into the middle of 1993. Many energy suppliers maintain accessible billing records only for the most recent 13 months. Thus, at the time of reporting, the data available did not cover the beginning of 1992.

- Bill coverage spanned less than a 1-year period.

- Bill coverage was for several sampled buildings combined. This occurred when no authorization form was obtained to authorize the supplier to provide data for individual buildings. In such cases, the supplier reported only annual totals for a group of sampled buildings summed together, using the electricity or natural gas worksheet.

- Bill coverage included nonsampled buildings or equipment outside the sampled buildings, as well as the one sampled building.

- Bill coverage excluded some of the building's occupants or tenants. This under coverage occurred when the energy supplier had several customers in a sampled building and was unable to identify all of them on the basis of the information provided by the Building Characteristics Survey respondent. In a few cases, energy suppliers were unwilling to release information on all customers in a building, even in aggregate form, without having a separate authorization from each.

- The problem of determining bill coverage was compounded by incomplete dates. In the most common case, the billing period date included a month and year, but not the day of the month.

To reconcile the discrepancies between the ideal billing data and what could actually be obtained, the following seven processing steps were taken:

1. Classify each set of bills, from a particular energy supplier for a particular building, as to coverage in terms of both building and time frame

2. Complete the billing dates for all bills

3. Annualize bills with full-year time frame coverage

4. Annualize bills with part-year time frame coverage

5. Adjust annualized bills, other than worksheet cases, for building over and undercoverage

6. Impute annual consumption and expenditures for buildings with completely missing data

7. Allocate worksheet totals among the buildings included on worksheets.

Each of these processing steps is explained below.

## Step 1. Classifying Coverage of Building and Time Frame

This classification was performed by the CBECS contractor as part of the data collection recordkeeping. To track responses to the mailed Energy Suppliers Survey, determination had to be made whether a response received represented complete data for a building. In many cases, follow-up letters converted initial responses from partial to complete, or more nearly complete. In other cases, the incomplete response was all that could be obtained.

### *Determining Time Frame*

An important aspect of the time-frame classification was determining why data were missing for part of calendar year 1992. The main question was whether consumption had actually taken place during the entire year or was actually zero during the unreported time.

If consumption occurred through the entire year, data might be missing for several reasons. One is that the supplier's active records might not go back far enough. Another is that data may simply have been lost from the supplier's record, even though in general these records did go back to the beginning of 1992.

A more complicated situation occurred when a new customer occupied a building in the middle of the target year. The data provided for this customer, for which the authorization form was signed, would be complete, but the data for the previous occupant, who consumed energy in the first part of the year, would be missing. In any case where part of the year's consumption data were missing, annual consumption would be understated if the reported 1992 data were treated as complete, rather than being inflated to account for the missing period.

The opposite situation could occur if a customer first occupied the building in the middle of the year, with no previous customer occupying the building. In this case, with no consumption during the first part of the year, annual consumption would be overstated if the reported data were annualized as if consumption occurred year round.

A special set of questions on the Energy Suppliers Survey forms was designed to determine if any change in customers had occurred during the target year, and if so how these customers were covered in the reported data. However, most suppliers did not answer these questions. As a general rule, data were treated as complete if they covered a full year, whether calendar 1992 or not. Part-year data were treated as incomplete, unless the supplier specifically indicated otherwise.

Particularly complicated were some electricity and natural gas cases where individual records were provided for each customer in a building with several customers. In most such cases, bills for all the customers covered the same time frame. Sometimes, though, different customers' records covered different time frames. In these cases, it was assumed that the data were complete for each customer, but the customers began or ended service at different times during the year. Aggregate consumption and expenditures were therefore computed for each time period by summing whichever customers had consumption during that period. If consumption was present for a particular customer in a particular period but expenditures were missing (or vice versa) aggregate expenditures (or consumption) were left as missing.

### *Determining Building Coverage*

Building coverage was determined from information obtained from both the Building Characteristics Survey respondent and the energy suppliers. Two types of problems could arise: (1) the energy bills covered more buildings than just the sampled building or (2) the energy bills omitted some of the building's occupants. In the first case, if the Building Characteristics Survey respondent indicated that a particular supplier's bill covered several buildings, the total square footage of buildings on that bill was requested. Then a disaggregation factor was computed as the ratio of the sampled building's square footage to this total square footage. In some cases, the supplier indicated that a bill covered additional, nonsampled buildings, though the Building Characteristics Survey respondent indicated otherwise. In these cases, the disaggregation ratio was computed using floorspace taken from listing information, or from the supplier's estimate. Disaggregation factors were always computed using the same source of information for both the total and the sampled building's floorspace: either the Building Characteristics Survey respondent, the listing information, or the supplier. Some suppliers, particularly for district heating and cooling, did not provide floorspace figures, but did give an estimate of what percentage of the reported consumption took place in the sampled building; these percentages were used directly as disaggregation factors.

When the information required to compute a disaggregation factor was unavailable from any source, a flag indicating that disaggregation was needed, but not possible, was placed on the building records. In these cases, annual consumption and expenditures were imputed as if the data for the building were completely missing.

In the second case, when the billing data omitted some customers in a building, an aggregation factor was computed. This factor was usually the ratio of the number of customers in the building to the number reported. Where more detailed information was available, the aggregation factor was the ratio of the total building floorspace to the floorspace occupied by the reported customers.

## Step 2. Complete Billing Dates

Virtually all missing billing dates were one of two types. The first type of dates that were incomplete had the month and year entered, but the day was missing for the beginning and ending dates of all billing periods on a record. These cases were imputed by assigning "16" to each beginning date and "15" to each ending date.

The second type of incomplete dates were missing the day of the month for some, but not all, billing periods. For each case of this second type, the billing periods affected were either bounded (surrounded by billing periods with known beginning and ending dates), or unbounded (either at the beginning or end of the set of billing periods). Any set of consecutive bounded billing periods with missing dates was assigned billing dates that would make all billing periods in the set have as close to the same number of days as possible. Unbounded billing periods were assigned beginning and/or ending dates as needed so that the number of days in each unbounded period was the same as the median number of days in billing periods of known length.

## Step 3. Annualizing Full-Year Data

One of the main reasons that the CBECS requested energy supplier data from December 1991 through January 1993 was to assure that 1992 consumption would be completely accounted for in the case of a complete response. However, unless a billing period happened to end on December 31, 1991, or December 31, 1992, consumption as reported by the energy suppliers ran over from the target period of calendar 1992, forward into 1993 and backward into 1991. In general, then, procedures were required to trim away these excess data. For this trimming, different procedures were used for continuous- and discrete-delivery energy sources.

For continuous-delivery energy sources (electricity, natural gas, and district sources), consumption and expenditures for a billing period extending into 1993 were adjusted by splitting the overlapping period into two subperiods, one running from the beginning date through December 31, the other from January 1 through the billing or meter reading date. Consumption and expenditures were prorated according to the number of days in each subperiod, and the consumption and expenditures for the subperiod that fell in 1992 were included in the total expenditures and consumption for 1992. An analogous procedure was used for a billing period extending into 1991. The assumption that the use of continuous-delivery energy sources took place at a constant rate throughout the billing period may be incorrect for any particular building. However, the procedure should yield approximately unbiased overall estimates.

Billing periods extending outside 1992 did not affect the discrete-delivery energy source (fuel oil) because, for this energy source, all deliveries during 1992 were accumulated. For fuel oil, the ending dates on the bills were used to determine which bills were for deliveries during 1992. No attempt was made to prorate bills, since there was no necessary connection between billing dates and consumption, as was the case for continuous-delivery energy sources.

For cases where the billing time frame covered a full year but was shifted so that either the beginning or the end of 1992 was not included, a similar procedure was used. In these cases, the data were annualized to a 1-year period within the reported time frame, overlapping as much as possible with 1992. The amount of shifting required to obtain 1-year periods is shown in Table B1 for electricity, natural gas, and district heat. A limited amount of shifting (involving 49 sampled buildings) was also performed for fuel oil.

## Step 4. Annualizing Part-Year Data

The annualization procedures for cases that had partial billing data, but less than a full year, were also different for continuous- and discrete-delivery energy sources. For continuous-delivery energy sources, the number of reported days of consumption was at least as large as the number of reported days of expenditures for almost all sets of bills. Thus, the major problem was to find methods of analyzing the incomplete consumption data. Expenditures were then annualized using the partial expenditures data and the annualized consumption data. The distributions of sampled buildings by days of reported consumption and expenditures data for continuous-delivery energy sources are given in Tables B2, B3, and B4.

## Table B1. Days of Data from Outside Calendar Year 1992 Used to Obtain Annual Estimates

| Shift of Reporting Period from Calendar Year 1992 | Number of Buildings | | | | | |
|---|---|---|---|---|---|---|
| | Sample | | | Population (thousand) | | |
| | Electricity | Natural Gas | District Heat | Electricity | Natural Gas | District Heat |
| All Buildings with Reported Data | 5,754 | 3,706 | 339 | 4,064 | 2,379 | 45 |
| Over 30 Days into 1991 | 16 | 25 | 9 | 13 | 19 | 2 |
| 30 or Fewer Days into 1991 | 152 | 168 | 35 | 115 | 118 | 1 |
| No Days Shifted | 5,049 | 3,152 | 287 | 3,600 | 2,034 | 40 |
| 30 or Fewer Days into 1993 | 267 | 146 | 4 | 169 | 82 | 1 |
| 31 to 90 Days into 1993 | 241 | 156 | 3 | 144 | 96 | 2 |
| 91 to 180 Days into 1993 | 44 | 38 | 0 | 15 | 21 | 0 |
| Over 180 Days into 1993 | 12 | 21 | 1 | 8 | 9 | 0 |

Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A through F, 1992 Commercial Buildings Energy Consumption Survey.

## Table B2. Days of Reported Consumption and Expenditures Data for Electricity

| Days of Reported Electricity Data | Consumption | | | Expenditures | | |
|---|---|---|---|---|---|---|
| | Sample Number of Cases | Population Number of Buildings (thousand) | Estimated Consumption (trillion Btu) | Sample Number of Cases | Population Number of Buildings (thousand) | Estimated Expenditures (million dollars) |
| All Buildings | 6,568 | 4,611 | 2,609 | 6,568 | 4,611 | 57,619 |
| Days of Electricity Data | | | | | | |
| 30 or Fewer Days | 813 | 546 | 396 | 824 | 550 | 9,323 |
| 31 to 330 Days | 118 | 84 | 27 | 132 | 87 | 919 |
| 331 to 365 Days | 87 | 53 | 27 | 94 | 56 | 846 |
| 366 Days | 5,197 | 3,689 | 1,986 | 5,165 | 3,678 | 42,586 |
| Worksheets | 353 | 240 | 173 | 353 | 240 | 3,945 |

Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A through F, 1992 Commercial Buildings Energy Consumption Survey.

## Table B3. Days of Reported Consumption and Expenditures Data for Natural Gas

| Days of Reported Electricity Data | Consumption | | | Expenditures | | |
|---|---|---|---|---|---|---|
| | Sample Number of Cases | Population Number of Buildings (thousand) | Estimated Consumption (trillion Btu) | Sample Number of Cases | Population Number of Buildings (thousand) | Estimated Expenditures (million dollars) |
| All Buildings | 4,152 | 2,657 | 2,487 | 4,152 | 2,657 | 10,679 |
| Days of Natural Gas Data | | | | | | |
| 30 or Fewer Days | 439 | 277 | 226 | 517 | 309 | 1,403 |
| 31 to 330 Days | 170 | 71 | 57 | 138 | 62 | 244 |
| 331 to 365 Days | 73 | 42 | 110 | 71 | 43 | 168 |
| 366 Days | 3,233 | 2,120 | 1,796 | 3,189 | 2,097 | 7,619 |
| Worksheets | 237 | 147 | 298 | 237 | 147 | 1,245 |

Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A through F, 1992 Commercial Buildings Energy Consumption Survey.

**Table B4. Days of Reported Consumption and Expenditures Data for District Heat**

| Days of Reported Electricity Data | Consumption | | | Expenditures | | |
|---|---|---|---|---|---|---|
| | Sample Number of Cases | Population Number of Buildings (thousand) | Estimated Consumption (trillion Btu) | Sample Number of Cases | Population Number of Buildings (thousand) | Estimated Expenditures (million dollars) |
| All Buildings | 558 | 95 | 435 | 558 | 95 | 2,901 |
| Days of District Heat Data | | | | | | |
| 30 or Fewer Days | 237 | 52 | 181 | 234 | 51 | 1,540 |
| 31 to 330 Days | 15 | 3 | 9 | 5 | 2 | 18 |
| 331 to 365 Days | 6 | 1 | 1 | 6 | 1 | 10 |
| 366 Days | 300 | 38 | 243 | 313 | 40 | 1,333 |

Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A through F, 1992 Commercial Buildings Energy Consumption Survey.

The part-year annualization method for the consumption of continuous-delivery energy sources depended on the number of days of reported consumption. If at least 331 days were reported, then consumption for the missing portion of the year was imputed by computing the average consumption per day for the adjacent billing period(s), then multiplying by the number of days of missing data. In certain cases, at least 331 days of consumption were reported, but 366 days of expenditures were reported[20]. In these cases, the missing consumption was computed using the average price for billing periods in which both consumption and expenditures were reported. Summing all reported and imputed consumption then yielded the total annual consumption.

Expenditure imputations were performed after completion of all imputations for partially missing consumption since (1) consumption data were usually more complete than expenditures data; and (2) given a value for consumption, the expenditures could be estimated without a great deal of difficulty.

As was true for consumption, the imputation procedure for missing continuous-delivery expenditures was determined by the number of days of reported data. If 30 or fewer days of expenditures were reported, then the expenditures were treated as completely missing. Otherwise, expenditures were imputed based on average prices within the set of bills for a given building. Using bills where both consumption and expenditures were reported, the consumption and the expenditures were summed. The average price was then calculated as the sum of the expenditures divided by the sum of the consumption. This average price was multiplied by the reported (or imputed) consumption to obtain the estimated expenditures.

For fuel oil, a discrete-delivery energy source, the billing dates are not linked to the time of consumption. Thus, the annualized data represent the total deliveries of fuel oil during the year. Furthermore, unlike continuous-delivery bills, discrete-delivery bills tend to be irregularly spaced. Gaps between bills could represent either missing data or periods during which no deliveries were required. The completeness of a set of bills was determined by relying on reports of suppliers. A set of bills was treated as complete if the supplier stated that the bills were complete for the year, and treated as missing otherwise, even if a partial set of bills was available. Table B5 shows the numbers of sampled buildings by the completeness of reported fuel oil data.

Buildings rarely had more than one supplier for a continuous-delivery energy source, such as electricity, but multiple suppliers for fuel oil occurred frequently. If data for one or more of several suppliers were missing, even though responding suppliers had reported all their 1992 deliveries, these buildings were also treated as if no data were available.

Imputations for both deliveries and expenditures made use of the observed price(s). An average price $P_x$, for each set of bills, was computed using the data from billing periods in which both consumption and expenditures were reported. If expenditures were missing, the expenditures were imputed as $P_x$ times the quantity delivered on date x. For missing deliveries, the reported expenditures were divided by $P_x$ to impute the amount delivered.

[20]Because 1992 was a leap year, all annualization calculations were based on 366 days.

**Table B5. Completeness of Reported Consumption and Expenditures Data for Fuel Oil**

| Completeness of Data | Consumption | | | Expenditures | | |
|---|---|---|---|---|---|---|
| | Sample Number of Cases | Population Number of Buildings (thousand) | Estimated Consumption (trillion Btu) | Sample Number of Cases | Population Number of Buildings (thousand) | Estimated Expenditures (million dollars) |
| All Buildings | 1,230 | 560 | 279 | 1,230 | 560 | 1,400 |
| Complete | 768 | 379 | 180 | 769 | 379 | 920 |
| Partial | 1 | 0 | 0 | 0 | 0 | 0 |
| Missing | 461 | 180 | 93 | 461 | 180 | 480 |

Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A through F, 1992 Commercial Buildings Energy Consumption Survey.

## Step 5. Adjusting for Building Over and Undercoverage (Other Than Worksheets)

The annualization procedures for full- and part-year data adjusted for inconsistent time-frame coverage. After the nonmissing consumption and expenditures data were annualized, the annual values were adjusted for building coverage. Where data were requested from the supplier for a single sampled building, but were provided only for a group of buildings including the sampled one, or were provided only for a portion of the building, the coverage adjustment was a simple multiplication of the annualized consumption and expenditures by the disaggregation or aggregation factor. As described above under Step 1, this factor was computed by the survey contractor directly on the basis of information received on the Building or Suppliers Survey.

## Step 6. Imputing for Completely Missing Consumption and Expenditures

In a significant fraction of cases, the energy supplier did not provide the consumption or expenditures data for some or all billing periods or deliveries in 1992. Reasons for missing data included energy supplier refusal; archived, lost, or destroyed billing records; and authorization form refusal on the part of the building respondent. In other cases, the energy supplier provided data, but either the building data were combined with those of nonsampled buildings and could not be disaggregated, or the consumption and/or expenditures were incomplete enough to be treated as missing.

The general approach taken to the problem of imputing annual consumption or expenditures was to annualize the complete or partial sets of bills first, then to use these annualized bills in regression equations to develop imputed values for the data that were totally missing. The regression imputation approach was chosen because data from the Building Characteristics Survey were already available for all of the buildings lacking energy supplier data. The first step was the estimation of missing consumption based on characteristics of buildings. After the consumption had been imputed, missing expenditures were estimated based on the reported or imputed consumption.

### *Completely Missing Consumption*

Each of the energy sources presented in this report was imputed separately, although the overall methodology was similar for all. The consumption imputation method is, therefore, described in general terms, referring to individual energy sources only where necessary. The regression equations were developed primarily to serve as adequate predictors of building consumption based on building characteristics. Simplicity and ease of estimation were also important considerations.

The data used to specify regression equations and estimate the regression parameters used for consumption imputation had to meet several criteria. Only cases with essentially complete consumption data were used. For continuous-delivery energy sources, "essentially complete data" included buildings with 331 to 366 days of reported consumption; for discrete-delivery energy sources, only buildings with completely reported deliveries were included. Any cases that were reported on forms with data from nonsampled buildings (or that lacked data for some customers within the sampled building) were eliminated if the disaggregation (or aggregation) factor (from Step 1) indicated that the sampled building accounted for less than half (or more than double) the total floorspace of all the buildings reported on the form. District heating cases were kept if the sampled building accounted for more than double or less than a tenth of the floorspace. Finally, any buildings with imputed values for building characteristics that were used as predictors in the regression equations were also eliminated.

The development of regression equations began by examining the distributions of the dependent variable, consumption. Because the distributions were found to be highly skewed a log transformation of the dependent variable was undertaken. Just as the consumption variable was highly skewed, so too were some of the potential regressor variables. Square footage, for instance, varied from 1,000 square feet to more than 1,000,000 square feet. Transformations of independent variables were evaluated by simple regressions of the log of consumption on various transformations of each potential quantitative variable. Plots of residuals versus predicted values from these simple regressions were also examined. As a result of these analyses, several key potential regressor variables, including the number of employees, square footage, and heated square footage, were also transformed to the log scale.

The principal activity within the building is an important determinant of consumption patterns. Therefore, for electricity, separate equations were developed for each of 13 principal building activities. For natural gas, which had a smaller sample size, 10 equations were developed. For fuel oil and district heat sample sizes were not large enough to permit regression equations to be fit by principal building activity.

The equations developed for the log of consumption were fit using ordinary least squares. Examination of residuals helped to isolate some reporting errors, but otherwise showed approximately normally distributed, homoscedastic residuals. However, the goal was to impute consumption, not the log of consumption. As an estimate of consumption, the back-transformed log prediction is a biased estimate.

Accordingly, the consumption values were calculated using parameter values estimated in two stages: the initial regression of log consumption on building characteristics, and a bias correction. The bias correction coefficient was estimated by (1) summing the total actual consumption of cases used to estimate the regression parameters, (2) summing the total of the back-transformed predicted values (from the log regression) for these same cases, and (3) dividing the sum of the actual values (1) by the sum of the back-transformed values (2).

### *Completely Missing Expenditures*

As for consumption, imputation for expenditures for each of the energy sources presented in this report was performed separately, although with a similar overall methodology. Again, the imputations are described in general terms, referring to individual energy sources only where necessary.

Energy supplier rate schedules are usually structured so that the price per unit of energy is lower as consumption increases. The rate schedule is usually a step function with the definition of steps and rates varying by energy supplier and by rate class. For the CBECS, rate schedules were not collected for the sampled buildings. Even the identity of the supplier was not disclosed to EIA. Therefore, a statistical procedure was needed to relate the expenditures to the consumption for imputation purposes.

As with the consumption imputations, the data used to specify the form and estimate the parameters of the expenditure imputation equations had to meet two criteria. First, only cases with essentially complete consumption and expenditures were used. For continuous-delivery energy sources, "essentially complete data" included buildings with 331 to 366 days of reported data for both consumption and expenditures; for discrete-delivery energy sources, only buildings with completely reported deliveries and expenditures were included. Any cases with data that were reported on forms with nonsampled buildings were eliminated if the disaggregation (or aggregation) factor (from Step 1) indicated that the sampled building accounted for less than half (or more than double) of the total floorspace of all the buildings reported on the form.

As a start, expenditures were plotted against consumption. Since both distributions were highly skewed, the log of expenditures was also plotted against the log of consumption. The latter set of plots disclosed a basically linear relationship between the log of expenditures and the log of consumption. The only noticeable departure from linearity was found at the low values of electricity and natural gas consumption, where the log of expenditures seemed to be unrelated to the log of the consumption. This cutoff apparently was due to base charges for these two energy sources, which dominated the total expenditures for low values of consumption. The breakpoint occurred at approximately 1,000 kWh for electricity and at approximately 10,000 cubic feet for natural gas. Therefore, buildings with annual consumption below these values were eliminated from the data used to fit the regression equations.

The approximately linear relationship observed between the log of expenditures and the log of consumption suggested an equation of the form:

$$\log(\text{expenditures}) = a + b \times \log(\text{consumption}).$$

This is for consumption above the cutoff. Transformed back from the log scale, this equation becomes:

$$\text{expenditures} = a \times \text{consumption}^{b}.$$

This equation expresses a plausible general relationship. If b equals one, then the parameter, a, can be interpreted as the price per unit consumed. If b is less than one, then the equation describes a situation in which the price per unit consumed declines with increasing consumption.

The above equation was estimated separately for metropolitan and nonmetropolitan areas within most Census division for electricity and natural gas. However, the CBECS sample size was insufficient to support this level of estimation for fuel oil, and district heat. For these two energy sources, the two parameters were estimated at the national level.

As was the case for consumption, the equations for the log of expenditures were fit using ordinary least squares. Transformation bias correction coefficients were also computed using the same procedure as for consumption.

If the reported or imputed value of electricity consumption for a building with missing expenditures was less than 1,000 kWh, then the expenditures were imputed as though the consumption were 1,000 kWh (the breakpoint identified in the plots of the log of expenditures versus the log of consumption). The same procedure was followed for natural gas, using a cutoff of 10,000 cubic feet for consumption. No cutoff was used for fuel oil or district heat.

## Step 7. Allocating Worksheet Totals

Worksheets combined consumption and expenditures for several sampled buildings and were used only for electricity and natural gas data. For each of these energy sources, the number of buildings with supplier data reported on worksheets represented about 5 percent of all sampled buildings supplied with the energy source.

The worksheet problem was not simply a matter of allocating an annual number among a set of buildings. In general, different reporting periods were given for each building on the worksheet, and the period lengths were rarely exactly 366 days long. In addition, the bills for a sampled building on a worksheet could include consumption in other, nonsampled, buildings just as was the case for sampled buildings not reported on worksheets.

A preliminary estimate of annual consumption and expenditures was made for each building on the worksheet using the regressions developed to impute completely missing data. A total for the set of cases on the worksheet was then estimated as:

$$\hat{W} = \sum_{i=1}^{n} \frac{\text{days}_i}{366} \times \frac{\hat{x}_i}{\text{adj}_i},$$

where

$\hat{W}$ = the regression-estimated worksheet total,

n = the number of buildings included on the worksheet,

$days_i$ = the number of days of data reported for the $i^{th}$ building,

$\hat{x}_i$ = the annual value estimated via regression for the $i^{th}$ building,

$adj_i$ = the aggregation/disaggregation adjustment for the $i^{th}$ building (as discussed in Step 1).

The ratio $\hat{x}_i/adj_i$ estimated the annual total that would have been reported for a building requiring aggregation or disaggregation by the factor $adj_i$. The ratio $days_i/366$ estimated the fraction (usually greater than one) of this annual total that would have appeared on the worksheet if $days_i$ of data were included for the building. The sum $\hat{W}$ was thus the regression-based estimate of what the worksheet total would have been.

The quantity (consumption or expenditures) for the $i^{th}$ building, $x_i$, was then calculated as:

$$x_i = \frac{W}{\hat{W}} \times \hat{x}_i,$$

where W was the supplier-reported worksheet total for the worksheet that included the $i^{th}$ building. The ratio $W/\hat{W}$ scaled the regression-imputed annual values, $\hat{x}_i$, to be consistent with the reported worksheet totals.

# Annual Peak Electricity Demand

Peak electricity demand data were requested for the same billing periods for which electricity consumption and expenditures were reported. (See Appendix G for copies of the electricity supplier forms.) Ideally, the metered demand represented the maximum consumption rate (in kW) during the billing period. However, two special data problems affect the availability of peak electricity demand data.

First, although virtually all electricity consumption is metered, peak electricity demand is metered where it is economical to do so. In general, peak demand meters are only installed for larger consumers of electricity. Second, in multicustomer buildings, each customer with a demand meter has its own peak demand. Since these peaks would rarely be coincident, the building peak cannot be taken as the sum of individual peaks. However, the overall building peak must be greater than or equal to the maximum customer peak.

Following Step 2 of "Annual Consumption and Expenditures," the peak electricity demand data was processed in three additional steps:

1. Using the billing data, each building was classified as either demand-metered or not demand-metered.

2. The annual peak demand, the season of the peak, and the annual load factor were determined for each building.

3. Peak demand and season of peak were imputed for demand-metered buildings missing these data.

These steps are described below.

## Step 1. Classification of Buildings

For the 1992 CBECS, a building was considered to be demand-metered if the billing data for any account within the building showed metered peak demand.[21]

[21]The 1989 CBECS obtained demand-metered information from both the building respondent and the energy supplier. However, there was considerable discrepancy between what the building respondent reported and the actual billing situation. As a result of the inability of the building respondent to adequately provide demand-metered data, the 1992 CBECS only obtained this information from the energy supplier.

## Step 2. Determination of Peak Demand and Related Items

For single-account buildings that were determined to be demand-metered in Step 1, the annual peak demand was taken as the maximum of the billing period peaks. For the few buildings that had part-year electricity billing data, the annual peak was taken as the maximum of the peaks in the reported billing periods. This approach results in a slight understatement of the annual peak, because the actual peak may have occurred during one of the unreported periods. However, since the number of buildings involved was relatively small, the difference between the part-year and full-year maxima would be small in most cases.

In multicustomer buildings, the overall building peak demand was not available. However, the overall peak had to be at least as high as the highest peak reported for any single customer. In buildings where one customer's peak was substantially larger than that of any other customer, that customer's peak would have been close to the overall peak. Therefore, in processing bills from multicustomer buildings, the peak demand for any single customer was designated as a "partial peak" (associated with part of the building electricity consumption), although the overall building peak was still treated as missing.

Before assigning the peak to a season, the month of the peak was found. Since the exact time of the billing period peak was unknown, the peak was taken to have occurred in whichever month contained the most days in the billing period during which the peak occurred. Peaks occurring November through April were then classified as winter peaks, while those occurring May through October were classified as summer peaks.

The annual load factor was then calculated, using previously calculated annual electricity consumption, as follows:

$$\text{annual load factor} = \frac{\text{annual consumption}}{366 \times 24 \times \text{peak annual demand}} .$$

As an edit, the annual load factor was calculated using the partial peak, and the partial peak was set to missing if the load factor was less than .10 or greater than 1.

## Step 3. Imputation for Missing Peak Demand

Although any electricity consumer has a peak demand, three types of buildings were missing peak demand:

- Buildings determined to be not demand-metered
- Buildings with completely missing supplier data
- Multicustomer buildings, and other buildings with partial peaks.

No attempt was made to impute for the first type of missing demand, mainly because buildings without demand-metering tended to be smaller than the demand-metered buildings, so that imputation would involve extrapolation beyond the range of the reported data. Accordingly, tables dealing with peak electricity demand have been limited to buildings with (reported or imputed) demand-metering.

Once the decision was made to exclude buildings that had not been demand-metered, imputation became a two-step process. First, it was necessary to impute whether the building with missing data was demand-metered. If the building was imputed to be a demand-metered building, then the peak and season of the peak were imputed. Table B6 shows the amount of each type of imputation that was necessary.

**Table B6. Item Response for Peak Electricity Demand Data**

| Response Category | Demand Metering | | Peak Demand | | Season of Peak | |
|---|---|---|---|---|---|---|
| | Sample Number of Cases | Population Number of Buildings (thousand) | Sample Number of Cases | Population Number of Buildings (thousand) | Sample Number of Cases | Population Number of Buildings (thousand) |
| Eligible Buildings . . . . . | 6,568 | 4,611 | 4,365 | 2,375 | 4,365 | 2,375 |
| Reported . . . . . | 5,428 | 3,846 | 3,171 | 1,786 | 3,050 | 1,752 |
| Imputed . . . . . | 1,140 | 765 | 1,194 | 589 | 1,315 | 623 |

Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A through F, 1992 Commercial Buildings Energy Consumption Survey.

Imputation of the demand-metering attribute made use of the relationship observed within suppliers between the presence of demand-metering and annual electricity consumption. Using buildings with reported data, the probability of being a demand-metered building was estimated as a logistic function of the annual consumption. The parameters estimated from the reported data regression were used to estimate probabilities for each unclassified building, and a uniform random number was generated. If the random number was less than or equal to the estimated probability, then the building was imputed to be demand-metered. For buildings imputed to be demand-metered, the season of peak demand was imputed by hot-decking, the same method used to impute missing items from the Building Characteristics Survey.

Finally, annual load factors were imputed for each building imputed to be demand metered. Values were imputed using parameters estimated from a linear regression of the logistic transformation of the annual load factor on various building characteristics (such as weekly operating hours, end uses of electricity, and percent of floorspace heated). Separate imputation equations were estimated for each of nine principal building activities. The imputed annual peak demand was then calculated by solving the load factor equation for the annual peak.

Load factors were imputed, and peak demand values calculated, for multiple-account buildings which had partial peaks (from Step 2). If the partial peak was less than the imputed peak, then the imputed peak was treated as the buildings' annual peak demand; otherwise, the partial peak was used.

Load factors and peak intensities were computed for each building reported or imputed to have metered demand. Also of interest are the analogous ratios over a utility service region, or other large area. The ratio of a region's consumption to the annual peak for the region as a whole would represent the average utilization of the region's generating capacity. The ratio of the region's annual peak to the total floorspace in the region would represent the average capacity requirement per square foot. However, the regional peak cannot be determined from the individual annual (or even monthly) peaks alone, since these peaks are not coincident. That is, the individual peaks occur at different times, so that the sum of the individual peaks can be considerably greater than the overall regional peak.

# Additional Data Notes

## Energy Sources Used--Building and Supplier Survey Estimates

As explained in Appendix A, "How the Survey Was Conducted," the CBECS was conducted in two stages. During the first stage, the building representative was asked which energy sources were used in the building during 1992. In the second stage, the energy suppliers, identified by the building representative during the first stage, were asked to provide consumption and expenditures data. In some cases, contacts with the energy suppliers revealed inaccuracies in the Building Characteristics Survey response as to which energy sources had been used in the building. All statistics in this report on energy sources used are based on the final determination made during the Energy Suppliers Survey.

When a supplier reported that a particular building was not a customer during 1992, calls were made to the building respondent to determine the reason for the discrepancy. In some cases, a different supplier was identified for the same energy source. In others, it turned out that the energy source had not actually been used; in some of these cases, a different energy source was identified instead. For example, natural gas may have been reported originally, but the callback determined that natural gas was consumed only in a central plant outside the sampled building, while the building itself used district steam, which had not been reported originally. In this case, natural gas would be coded as "not used in the building," and district steam would be added as "used in the building." The net discrepancies between the Building Characteristics Survey and Energy Suppliers Survey estimates for the use of each energy source were small for both the building counts and the floorspace totals (Tables B7 and B8).

The Energy Suppliers Survey was able to correct the energy sources used, only in cases where a supplier had been misreported as supplying a particular building with an energy source. If the Building Characteristics Survey respondent happened to omit an energy supplier, but reported all the other supplier data correctly, the omitted supplier would not have been discovered. However, the number of such cases was probably quite small.

In some cases, a supplier reported that a particular building had been a customer for a given energy source, but not during calendar year 1992. For continuous-delivery energy sources (electricity, natural gas, and district heating and cooling), the building was classified as not using the energy source. For the discrete-delivery energy source fuel oil, though, the building was classed as using the energy source, but with zero consumption and expenditures for 1992. Thus, for example, those buildings whose respondents reported that fuel oil was used during 1992, but which received no fuel oil deliveries in that year, were included in the count of buildings and floorspace using fuel oil, though they did not contribute to the fuel oil delivery total.

The revised information on the type of energy sources that were used in the building had an impact on the energy end-use data also. The Building Characteristics Survey data on the type of energy sources that were used for a particular end use were collected in concert with the data on energy sources used. (See Appendix G for copies of the survey forms.) Edit checks on the Building Characteristics Survey data assured consistency between energy sources reported for end uses and energy sources reported at all. However, when the information on energy sources used "at all" was revised during the Energy Suppliers Survey, no new information was obtained on energy sources used for particular end uses. As a result, some energy sources were dropped from a building's list of energy sources used, even though these energy sources had end uses reported. Conversely, no associated end uses were coded for energy sources that were added for a building. For any energy source whose use was changed from "yes" to "no" for a particular building, the use of that energy source for any given end use was also changed to "no." However, the end use was still treated as having been performed in the building. That is, it was assumed that the building respondent correctly reported, which end uses were performed, even if the energy source used for the end use had been incorrectly reported. This approach left some buildings identified as having a particular end use, but with no energy source indicated for that use.

### Natural Gas Transported for the Account of Others

The 1992 CBECS attempted to collect data on natural gas transported for the account of others[22] from both the building respondent and the natural gas suppliers—both utility suppliers and nonutility suppliers. Natural gas transported for the account of others is a type of purchasing arrangement where large natural gas users purchase their natural gas directly from a source other than the local distributing company (LDC) or utility. The LDC then delivers the gas to the building via the local pipelines.

Schedule A of Form EIA-871C-1a requested: (1) consumption and expenditures for gas bought from the LDC; (2) consumption of gas purchased other than from LDC; (3) delivery charges for gas purchased from other than the LDC; and (4) total charges for this gas (See Appendix G, "Survey Forms").

[22]"Gas transported for the account of others" is also referred to as "direct purchase gas," "spot market gas" or "transportation gas."

**Table B7. Energy Sources Used As Reported on Building Questionnaire and Energy Supplier Survey, Number of Buildings**
(Thousand)

| Reported Use | Energy Sources | | | | | |
|---|---|---|---|---|---|---|
| | Electricity | Natural Gas | Fuel Oil | District Steam | District Hot Water | District Chilled Water |
| **Total Reported on Building Questionnaire** | 4,616 | 2,665 | 559 | 64 | 39 | 28 |
| Unchanged Based on Energy Supplier Survey | 4,611 | 2,655 | 558 | 61 | 38 | 28 |
| Deleted Based on Energy Supplier Survey | 4 | 10 | 1 | 2 | 1 | 0 |
| Added Based on Energy Supplier Survey | NC | 2 | 2 | 0 | 1 | NC |
| **Total Based on Energy Supplier Survey** (Final Resolution) | 4,611 | 2,657 | 560 | 62 | 39 | 28 |
| **Not Used Based on Both Building and Energy Supplier Survey** | 190 | 2,138 | 4,245 | 4,742 | 4,766 | 4,777 |

NC No cases in responding sample.
Notes: • See the "Glossary" for explanation of abbreviations and definitions of terms used in this report. • Items may not add due to rounding.
Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A through F, 1992 Commercial Buildings Energy Consumption Survey.

**Table B8. Energy Sources Used As Reported on Building Questionnaire and Energy Supplier Survey, Floorspace**
(Million Square Feet)

| Reported Use | Energy Sources | | | | | |
|---|---|---|---|---|---|---|
| | Electricity | Natural Gas | Fuel Oil | District Steam | District Hot Water | District Chilled Water |
| **Total Reported on Building Questionnaire** | 66,549 | 45,097 | 13,218 | 4,571 | 1,310 | 2,066 |
| Unchanged Based on Energy Supplier Survey | 66,525 | 44,975 | 13,204 | 4,466 | 1,068 | 1,914 |
| Deleted Based on Energy Supplier Survey | 25 | 121 | 14 | 105 | 241 | 152 |
| Added Based on Energy Supplier Survey | NC | 18 | 10 | 7 | 12 | NC |
| **Total Based on Energy Supplier Survey** (Final Resolution) | 66,525 | 44,994 | 13,215 | 4,473 | 1,080 | 1,914 |
| **Not Used Based on Both Building and Energy Supplier Survey** | 1,327 | 22,761 | 54,648 | 63,298 | 66,555 | 65,810 |

NC No cases in responding survey.
Note: See the "Glossary" for explanation of abbreviations and definitions of terms used in this report.
Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A through F, 1992 Commercial Buildings Energy Consumption Survey.

Analysis of the natural gas transported for the account of others data collected in the 1989 CBECS indicated that while the LDC could report the volume of natural gas used, they often could not report the total expenditures, since the LDC did not know the purchase price the building paid for the independent purchase of this gas. Consequently, in the 1992 CBECS, the building respondent was asked to provide the expenditure information such as wellhead costs, city gate price, LDC charge and other costs associated with this type of purchasing arrangement. However, this proved to be an area where the building respondent was unable to provide the requested information, so expenditure data for natural gas transported for the account of others was taken from the supplier forms.

Since local distribution companies know the total volume of natural gas delivered, the total consumption data seem complete. (If natural gas consumption was completely missing, then the volume was imputed as described in Step 6 of "Annual Consumption and Expenditures"). The allocation of consumption between transported gas and local utility-owned gas was then imputed by hot-decking the proportion of gas that was transported gas. This method allowed imputed buildings to have both transported and local utility gas, as might happen if (1) building demand exceeded the direct purchase contract amount or (2) the building switched to or from a direct purchase contract during the year.

This report contains estimates of the number of buildings, floorspace, total natural gas consumption, and consumption of natural gas transported for the account of others (Table 3.40). Table 3.40 also includes the percentage of natural gas volume which was natural gas transported for the account of others. Overall, 15 percent of natural gas consumed in commercial buildings was gas transported for the account of others. This figure is very close to the amount reported in the *Natural Gas Annual 1992*, where 17 percent of commercial natural gas deliveries in 1992 were estimated to be for the account of others.

Estimating consumption and expenditures could become complicated because frequently the LDC filled out the gas transported for the account of others portion of the supplier form since they knew that the gas being provided was transportation gas. Conversely, gas companies which provide only natural gas transported for the account of others did not always fill in the form correctly. They often filled in the first available space, which was intended for utility gas only. Similar confusion occurred when filling in transported gas expenditures. The LDC would be expected to fill out the transport charges column but because this was the only expense collected by the LDC, they sometimes filled it in the "total" column. Finally, since the same volume of gas was reported by the LDC and the transportation gas company, double reporting of volumes sometimes occurred. All these problems were identified by visual inspection of the appropriate records.

### Demand-Side Management Participation

The data on DSM participation during the three years prior to the 1992 CBECS that are presented in Section 3, "Detailed Tables," of this report (with the exception of Table 3.49) and in the companion volume to this report, *Commercial Buildings Characteristics 1992*, are based solely on information gathered from the building questionnaire. The DSM data in Section 2, "At a Glance," on the other hand, is based on information gathered from the supplier survey. It would be expected that the information from the building questionnaire would indicate higher levels of DSM participation, because the data in the tables do not restrict participation to DSM programs sponsored by an electric utility, and include participation regardless of the sponsor of the program. However, the data in Section 2 indicate much higher rates of participation, even though they include only data from electric utilities. The data in this section have been imputed for nonresponse, but that is not the source of the discrepancy. The statistics in Tables B9 and B10 have been categorized to separate the programs sponsored by electric utilities from programs sponsored by natural gas utilities (as indicated by the utilities themselves and the building respondent), and cases in which the question was not answered have been included. From these tables it can be seen that suppliers consistently indicated a higher rate of participation than building respondents. This is perhaps because the suppliers had more detailed records of DSM participation, and also because the utilities may have included the distribution of general information, such as brochures about their programs, as a type of DSM participation.

### Table B9. Commercial Buildings Participating in Electric DSM Programs

| | Participation According Building Questionnaire | | | |
|---|---|---|---|---|
| | Electricity Not Used | Yes | No | Not Ascertained |
| **Participation According to Supplier Survey** | | | | |
| Electricity Not Used | 177 | 0 | 4 | 2 |
| Yes | 0 | 327 | 1,504 | 113 |
| No | 0 | 154 | 3,014 | 206 |
| Not Ascertained | 0 | 91 | 932 | 227 |

Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A, EIA-871E-1b, 1992 Commercial Buildings Energy Consumption Survey

**Table B10. Commercial Buildings Participating in Natural Gas DSM Programs**

| | Participation According Building Questionnaire | | | |
|---|---|---|---|---|
| | Natural Gas Not Used | Yes | No | Not Ascertained |
| **Participation According to Supplier Survey** | | | | |
| Natural Gas Not Used | 2,582 | 0 | 13 | 4 |
| Yes | 1 | 11 | 284 | 29 |
| No | 4 | 38 | 2,528 | 195 |
| Not Ascertained | 0 | 20 | 894 | 148 |

Source: Energy Information Administration, Office of Energy Markets and End Use, Forms EIA-871A, EIA-871C-1b, 1992 Commercial Buildings Energy Consumption Survey.

**Comparison of CBECS and Form EIA-861 Data:** Of the electric utilities surveyed by CBECS, 46 percent of them reported that they had a DSM program, 40 percent reported that they did not have a DSM program, and 13 percent did not respond. These figures were compared to the data reported on Form EIA-861, "Annual Electric Utility Report," which is utility-reported data.[23] The findings corresponded very closely. Of the utilities that reported data on the Form EIA-861 and who were also in the CBECS sample, 50 percent reported that they had a DSM program and 50 percent reported that they did not have a DSM program. However, the utilities did not always have the same response to both the CBECS and the Form EIA-861. Of the utilities that responded to both surveys, 31 percent had inconsistent answers.

It was decided to use the responses to Form EIA-861 to override the CBECS responses where they conflicted for the following two reasons: (1) Form EIA-861 is mandatory, whereas CBECS had some nonresponse, and (2) there were four utilities that responded to the CBECS for multiple buildings but were not consistent across all of the buildings.

## Primary Energy Consumption for Electricity

The CBECS collects data on the amount of energy delivered to commercial buildings, the "site energy consumption." It does not collect data on the amount of energy needed to produce the site energy, the "primary energy consumption." However, concern with improving energy efficiency has promoted awareness of the need to account for the amount of energy lost in the production of the site energy, especially during the generation of electricity.

In the generation of electricity, large amounts of energy losses occur:

- When heat is converted into mechanical energy for turning electric generators
- When the power plant uses electricity for such uses as pumping water into elevated reservoirs in pumped-storage hydroelectric plants
- When electricity is transmitted and distributed from the power plant to the consumer.[24]

Measuring the amount of these energy losses is complicated because their amount varies from year to year and from utility plant to utility plant, depending on the conversion process and energy sources used, the particular mix of energy sources, and the efficiency of the utility plant. Since collecting data on these factors for each utility plant is obviously unreasonable within the framework of EIA consumption surveys, the amount of energy consumed to produce the electricity consumed on site in any given year can only be estimated. EIA bases this estimate on the approximate annual amount of fossil fuels (coal, natural gas, and petroleum products) used by steam-electric generating plants, which generate most of the Nation's electricity.[25]

[23]The number of electric utilities that sponsor a DSM program is reported on Form EIA-861, "Annual Electric Utility Report," and collected by EIA, Office of Coal, Nuclear, Electric and Alternate Fuels.

[24]Although energy losses also occur during the production of natural gas, fuel oil, and district heat, they are so small compared with those occurring during the production of electricity that they are not considered in measuring primary energy consumption in this report.

[25]The fossil fuels, especially coal, provide the principal energy sources for the generation of electricity. Nuclear and hydroelectric power are used to a lesser extend, with wood/waste, wind, geothermal, and solar energy supplying only a small amount of energy for electricity generation.

In 1992, U.S. steam-electric utility plants are estimated to have used approximately 10,302 Btu of fossil-fuel energy to generate 1 kilowatthour of electricity—or approximately 3.02 Btu of fossil-fuel energy to generate 1 Btu of electricity, since 3,412 Btu equals 1 kilowatthour of electricity.[26] Accordingly, in this report:

- Estimates of site electricity consumption in kilowatthours can be converted to estimates of primary energy consumption by using 10,302 as the conversion factor.

- Estimates of site electricity consumption in Btu can be converted to estimates of primary energy consumption by using 3.02 (10,302 divided by 3,412) as the conversion factor.

Estimates of primary energy consumption for electricity using a particular conversion factor should be considered as rough estimates only, but they do provide a more comprehensive picture of the amount of energy used in the commercial sector in a given year than is gained by merely looking at site consumption.

## Estimation of Sampling Errors

Sampling error, as described in the introduction to this appendix, is the random difference between the survey estimate and the true population value. This difference arises because a random subset, rather than the whole population, is observed. The typical magnitude of the sampling error is measured by the standard error of the estimate. The standard error is the root-mean-square difference between the estimate based on a particular sample and the value that would be obtained by averaging estimates over all possible samples.

If the estimates are unbiased, meaning there is no systematic error, this average over all possible samples is the true population value. In this case, the standard error is simply the root-mean-square difference between the survey estimate and the true population value. If systematic error is present, however, this bias is not included in the error measured by the standard error. Thus, the standard error tends to understate the total estimation error if there are noneligible biases.

In principle, random errors can be contributed to the estimate by sources other than the sampling process. Such additional sources of random error include random errors by respondents and by data entry staff, and random unit nonresponse. To recognize these additional sources of variation, the definition of the sampling process can be expanded to include not just the selection of buildings but all steps required to obtain a set of responses. Under this expanded definition, all random errors can be regarded as sampling errors. The procedures designed to estimate the sampling error must, therefore, incorporate all random components of the estimation process.

### Jackknife Replication

Throughout this report, standard errors are given as percents of their estimated values, that is, as RSE's. Computations of standard errors are more conveniently described, however, in terms of the estimation variance, which is the square of the standard error.

For some types of surveys, a convenient algebraic formula for computing variances can be obtained. However, the CBECS used a list-supplemented, multistage area sample design (see Appendix A, "How the Survey Was Conducted") of such complexity that it is virtually impossible to construct an exact algebraic expression for estimating variances. In particular, convenient formulas based on an assumption of simple random sampling, typical of most standard statistical packages, are entirely inappropriate for the CBECS estimates. Such formulas tend to give severely understated standard errors, making the estimates appear much more accurate than is the case.

The method used to estimate sampling variances for this survey was a jackknife replication method. The idea behind replication methods is to form several pseudoreplicates of the sample by selecting subsets of the full sample. The subsets are selected in such a way that the observed variance of estimates based on the different pseudoreplicates estimates the sampling variance in the overall estimate.

[26]"Table A8. Approximate Heat Rates for Electricity," *Monthly Energy Review* (August 1994), p. 165.

The replication method used begins by pairing first-stage sampling units, such that the units in each group represent two or more independent draws from the same pool of first-stage units, and draws for different groups are also independent. This grouping of first-stage sampling units must be done in accordance with the way the sampling was actually conducted. For the 1992 CBECS, 44 groups of first-stage sampling units were created in this way.

The $k^{th}$ jackknife pseudoreplicate sample set is obtained by deleting all observations from one of the members in the $k^{th}$ group and multiplying the weights on all cases in the other group member by 2 if there are 2 members in the group and by 1.5 if there are 3 members in the group. Observations in all other groups are unaffected. The $k^{th}$ pseudoestimate is then obtained from this pseudoreplicate sample by following all the steps used to construct the full-sample estimate.

The variances are estimated from the pseudoestimates in the following way. Let $X'$ be a survey estimate (based on the full sample) of characteristic X for a certain category of buildings. For example, X may be the total square footage of buildings using natural gas in the Midwest. Let $X_k'$ be the pseudoestimate of X based on the $k^{th}$ pseudoreplicate sample. The estimated variance of the full-sample estimate $X'$ is then given by:

$$S_{X'}^2 = \sum_{k=1}^{44} (X_k' - X')^2 .$$

The standard error of $X'$ is given by:

The relative standard error (percent) of $X'$ is obtained from this standard error as:

$$RSE_{X'} = \left( \frac{S_{X'}}{X'} \right) \times 100 .$$

### Effects of Missing Data on Error Estimation

Earlier in this appendix, the procedures used to adjust for unit and item nonresponse were described. Because the missing cases and the responding cases used to adjust for them arise randomly (within adjustment groups) nonresponse contributes to the estimation variance, even when appropriate adjustment procedures are used to remove the nonresponse bias. Half-sample replication estimates of variance account for this component of variance only if adjustments are made separately for each replicate.

To capture the effect of random nonresponse on the variance of estimates, a separate unit nonresponse adjustment factor, as described in the section on "Unit Nonresponse Adjustment," was computed for each pseudoreplicate. Thus, each replicate estimate was computed using a slightly different set of adjusted weights.

As in previous surveys, RSE's of consumption, expenditures, and peak-demand related items were computed excluding cases that were imputed by regression. RSE's of consumption and expenditures for the sum of major fuels were computed excluding cases where more than half of the quantity had been imputed by regression. The practice of eliminating imputed values was supported by a nonresponse simulation study, which found the resulting RSE estimates to be reasonable approximations to the true RSE's.[27] However, basing estimated RSE's on reported cases is an ad hoc procedure. This practice may be misleading, especially for fuel oil and district heat, where a substantial portion of the estimated totals were imputed (see Tables B4 and B5).

[27]E. M. Burns, Imputing for Missing Survey Data: "Energy Consumption in Commercial Buildings," Proceedings of the Business and Economic Statistics Section of the American Statistical Association (1987).

### Generalized Variances

For every estimate in this report, the RSE was computed by the methods described above. This was the RSE used for any statistical tests or confidence intervals given in the text, or to determine if the estimate was too inaccurate to publish (RSE greater than 50 percent).

Space limitations prevent publishing the complete set of RSE's with this document. Instead, a generalized variance technique is provided, by which the reader can compute an approximate RSE for each of the estimates in the main summary tables. For an estimate in the $i^{th}$ row and $j^{th}$ column of a particular table, the approximate RSE is given by the simple formula:

$$RSE_{i,j} = R_i C_j,$$

where $R_i$ is the RSE row factor given in the last column of row i, and $C_j$ is the RSE column factor given at the top of column j. (See Section 3, "Detailed Tables," for a discussion of how to use the RSE Row and Column factors in this report.)

### Derivation of Row and Column Factors

The row and column factors are determined from a two-factor analysis of the table of RSE's, on the basis of the model:

$$\log(RSE_{i,j}) = m + a_i + b_j .$$

The least-squares estimates for this model are given by:

$$m = \overline{\log(RSE)}$$

$$a_i = \overline{\log(RSE_{i.})} - \overline{\log(RSE)}$$

$$b_j = \overline{\log(RSE_{.j})} - \overline{\log(RSE)} ,$$

where $\overline{\log(RSE)}$ is the mean of $\log(RSE_{i,j})$ over all rows i and columns j, $\overline{\log(RSE_{i.})}$ is the mean over all columns j for a particular row i, and $\overline{\log(RSE_{.j})}$ is the mean over all rows i for a particular column j. The row and column RSE factors are then computed as:

$$R_i = \log^{-1}(m + a_i) = \log^{-1}( \overline{\log(RSE_{i.})} )$$

$$C_j = \log^{-1}(b_j) = \log^{-1}( \overline{\log(RSE_{.j})} - \overline{\log(RSE)} ) .$$

The RSE row factor, $R_i$, is thus the geometric mean of the RSE's in row i, and the RSE column factor, $C_j$, is an adjustment factor with geometric mean equal to 1.0.

For a few table cells, there were no sample cases, hence, no estimate and no RSE. As a result, some of the arrays of direct estimates $RSE_{i,j}$ had a few missing values. In such cases, the formulas given above for row and column factors still apply, but only after appropriate estimates have been substituted for the missing values.[4] In cases where a statistic was not publishable, because of a high RSE or small cell sample size, the value of $RSE_{i,j}$ was set to missing, so that the computed row and column factors are based only on published cases.