Appendix C

Nonsampling and Sampling Errors

Appendix C

Nonsampling and Sampling Errors

Introduction

All of the statistics published in this report are estimates of population values, such as the total floorspace of commercial buildings in the United States. These estimates are based on reported data 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 value errors 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 two sections that follow this introduction, "Data Collection Problems" and "Nonresponse," 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); and (3) nonresponse on a particular question from a responding building (item nonresponse). The section "Data Collection Problems" addresses some of the difficulties encountered in trying to obtain meaningful data on questionnaire items in the 1992 survey. The section "Nonresponse" presents in detail the procedures used to handle unit and item nonresponse.

Most unit nonresponse cases were caused by a respondent's unavailability or refusal to participate in the survey. Item nonresponse resulted when the building 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 of all stages of survey design and field procedures. The wording and format of survey questionnaires, the procedures used to select and train interviewers, 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 B, "How the Survey Was Conducted."

Sampling errors, on the other hand, are random differences between the survey estimate and the population value, that occur because the survey estimate is calculated from a randomly chosen subset of the entire population. 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. However, the sample design permits sampling errors to be estimated. The section, "Estimation of Standard Errors," describes how the sampling error is estimated and presented for statistics given in this report.

Nonsampling Error

Data Collection Problems

Even though the interviewer was instructed to conduct the interview with the person most knowledgeable about the building, there was a great deal of variation in how much CBECS respondents knew about their buildings. Some respondents did not know some of the information requested; some were able to provide certain information only if the questions were expressed in the particular terms they understood. This presented a special challenge to the CBECS questionnaire designers: with such a diverse population of respondents, it is difficult to construct standard wording for energy concepts that would be understood by all respondents. (See Appendix G, "Survey Forms," for a copy of the Buildings Questionnaire.) Additionally, even when a question is worded clearly and the respondent understands the question and has the required knowledge, simple clerical errors (possibly the fault of the questionnaire layout) can sometimes lead to inaccuracies in the data.

Following is a summary of some difficulties that EIA staff has identified with the survey responses. The extent of these comments should not be viewed as a failure of the questionnaire or the interview process; the data collection process worked well. Rather, these comments indicate areas that require further refinements to improve overall data quality.

Principal Building Activity

The principal building activity refers to the primary function or activity that occupies the most floorspace in the building sampled. In some cases, particularly if the sampled building was one of a number of buildings on a facility, the respondent reported the overall function of the facility or establishment to which the building belonged. In CBECS, for instance, a library is classified as a public assembly building, but a library on a university campus may have been reported as an education building (academic or technical instruction). To help alleviate this confusion, the 1992 CBECS asked a separate question for the overall facility activity for those buildings identified as being part of a facility. The principal activities of respondent buildings were checked against other available information, including the facility activity, interviewer observations, and the building's name, and recoded if an obvious assignment error was made.

Another difficulty with identifying principal building activities is that buildings with the same title may, in fact, have different primary functions. For example, space in a building referred to as a "courthouse" can be devoted primarily to office activities (office), to jail cells (public order and safety), or to hearing rooms (public assembly).

For some buildings, no one activity occupied 50 percent or more of the floorspace, but the activity occupying more space than any other was either industrial or residential. For example, it is possible for a building to have 30 percent of the floorspace devoted to assembly, 30 percent to food sales, and 40 percent to residential. Since more than 50 percent of the floorspace was occupied by commercial activity, these buildings were retained in the sample as commercial buildings, but were included in the "Other" category.

Operating Hours

During the imputation phase of the survey, it became apparent that there were some buildings with anomalous operating hours, which warranted a closer investigation of operating hours. For example, some a.m. times had been reported as p.m. times, and vice versa, apparently through an interviewer or respondent error. Other cases were apparently reported accurately--some buildings do indeed have unusual operating hours.

In 1992, as in 1989, data on operating hours were not ascertained in cases where the building respondent reported that the building was not in use during the previous 12 months. These cases are treated in the detailed tables as having zero operating hours per week. This represents a change from the 1986 survey questionnaire.

In 1992, operating hours were also determined for each day of the week, unlike 1989, which asked for the operating hours for the category "Monday through Friday" and then separately for "Saturday" and for "Sunday." This change allowed for more accurate validation of the total operating hours for the building.

Number of Workers

The CBECS collects data on the number of people who work in commercial buildings. Included in this number are volunteer workers, but not clients, students, or employees who work away from the building. A change in the question between the 1986 and the 1989 CBECS resulted in a somewhat smaller estimate of employment totals for 1989 than the corresponding estimates for 1986. The 1986 CBECS asked for the total number of people working in the building across all shifts. Although this was not obvious in the 1986 questionnaire, it was specified in the interviewer instructions. While it is not inconceivable that some respondents in 1986 may have given the number of workers for the main shift, the responses are, for the most part, consistent with the total number working across all shifts. On the other hand, the 1989 survey specifically asked for the number working during the main shift. The total number of people who work in the building provides a better basis for estimating floorspace by region from employment data, which tend to be more readily available from economic series. The number working during the main shift gives a more meaningful number with to estimate the capacity of the building's energy-using systems. In order to compare the 1992 CBECS number of workers with both the 1986 and 1989 CBECS, the 1992 CBECS asked both the total number of workers across all shifts and the number of workers for the main shift.

In the 1992 CBECS, if a building was not in use during the previous 12 months it was still included in the less- than five category of number of workers.

Heating and Cooling

The phrasing of questions on heating and cooling equipment has presented difficulties in every CBECS conducted thus far and, unfortunately, illustrates difficulties both in question wording and in respondent knowledge. Commercial buildings' heating and cooling systems vary greatly in design and complexity. The CBECS questionnaire designers try to formulate a few questions that could broadly characterize a building's heating and cooling system.

In previous CBECS, some building respondents (especially those from larger buildings), found the questions to be too general to adequately describe their buildings' systems. Other building respondents lacked even the rudimentary knowledge of their buildings' systems required by the questionnaire. To alleviate some of the problems encountered in earlier CBECS in which inconsistencies appeared between types of equipment, fuel sources and the distribution system, the 1992 CBECS questionnaire limited the respondents' choices in such a way that only sensible combinations of heating or cooling equipment with distribution equipment could appear. Additionally, a general question was added to the questionnaire, which asked the respondent to describe the heating and cooling system. This verbatim description was not coded on the computer file, but was of immeasurable value in deciphering the respondents' intentions.

The question of whether the building used "heat pumps" also confused a number of respondents. Two types of problems were associated with the use of heat pumps. First, 134 respondents indicated that they used a heat pump for either heating or cooling but not for both heating and cooling. This may have resulted because the placement of the heat pump category in the cooling question was different from the heating question. (See Appendix G, "Survey Forms," for a copy of the Buildings Questionnaire.)

The second problem pertaining to heat pumps was more troublesome. Some respondents indicated that they used heat pumps for heating but they listed only natural gas as their heating fuel. To date, there are only prototypes of natural gas heat pumps. After further investigation, the respondents that listed heat pumps as heating equipment had been mistaken. The heat pumps were most often confused with packaged units.

Gas Transported for the Account of Others

For the first time, the 1992 CBECS building respondents were asked whether they purchased natural gas directly from a source other than the local distributing company (LDC). This purchasing arrangement is known as "gas transported for the account of others." It is also known as "direct purchase gas" or "spot market gas." The 1992 CBECS data show that the larger buildings tend to be the ones that receive direct purchase gas.

In the 1989 CBECS, this information was asked of the energy suppliers only. Although suppliers could provide the volume of natural gas delivered they could not, in many cases, report the expenditures since they did not know the purchase price of the transported gas. It was believed that the building respondent would be better able to provide information about whether they purchased natural gas under this arrangement, who the suppliers were and what were the wellhead costs, city gate price, LDC charge, and other costs associated with gas transported for the account of others. This, however, proved to be another area where the building respondent had difficulty providing information. Of those reporting that they did buy natural gas under this purchasing arrangement, only 18 percent could report one or more of the costs associated with the purchase.

It appears that CBECS respondents, the people who are supposed to be most knowledgeable about the energy-using systems of the buildings are not the most knowledgeable about billing arrangements. In future CBECS, it may be necessary to target the person most knowledgeable about billing with a separate data collection effort in order to make reliable estimates about gas transported for the account of others.

Renewable Energy Sources

The CBECS attempted to collect information on the use of renewable energy sources in 1992 by including wood, photovoltaic cells (PVC's), and solar thermal panels in the list of possible energy sources that were used to supply energy to the building. An additional question was also asked about the use of special energy technologies, which included passive solar features, geothermal energy, and wind generation. Wood was used in about 2 percent of the buildings as an energy source. Table C1 shows the number of sample buildings reporting the use of various renewable energy sources and special energy technologies such as solar thermal panels, photovoltaic cells, passive solar features, geothermal energy and wind generation. The small number of respondents (less than 20 buildings) prohibited publishing the data in the detailed tables.

Table C1: Number of Sample Buildings Using Renewable Energy Sources and Special Energy Technologies, 1992

Renewable Energy Sources	Sample Cases
Total	6,751
Wood	74
Photovoltaic Cells	1
Solar Thermal Panels	8
Passive Solar Features	49
Thermal Energy Storage	53
Geothermal Energy	2
Well Water for Cooling	43
Waste Incineration to Produce Energy	19
Wind Generation	0

Source: Energy Information Administration, Office of Energy Markets and End Use, 1992 Commercial Buildings Energy Consumption Survey.

Nonresponse

Unit Nonresponse

The response rate for the 1992 CBECS, reported in Appendix B, was 91.1 percent. That is, of the 7,282 buildings eligible for interview, 8.9 percent did not participate in the Building Characteristics Survey. This rate was similar to that for the 1986 and 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. The method of estimation used was to calculate 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 rough size of the building, Census region, and metropolitan versus nonmetropolitan location.

Item Nonresponse

Table C2 contains item nonresponse rates for some of the building characteristics presented in this report. "Eligible" in this context refers to interviewed buildings to which the question item applied; certain sequences of responses to previous questions would make some question items not applicable for some respondents.

Nonresponses to several items in otherwise completed questionnaires 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. What characteristics 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 response for the particular item in the 1986 CBECS for those buildings that were surveyed in 1986) were used for specific items. 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.

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 thus could not be interviewed. Because these buildings were clustered in a few areas, they were not adequately represented by buildings elsewhere, and thus it was decided that the unit nonresponse adjustment procedure would not be the optimal way to compensate for these buildings. Instead, in those areas, it was decided to impute for all of the building characteristics, based on information available from the 1992 sample listing stage and from the 1986 survey. These imputations are included in the item nonresponse rates given in Table C2.

Estimation of Standard Errors

Sampling error, as described in the introduction to this appendix, is the difference between the survey estimate and the true population value due to using a random sample to estimate for a population. 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 non-negligible biases.

Table C2. Item Nonresponse Percentages for Selected Building Characteristics

Building Characteristics	Eligible Buildings	Number Missing	Percent Nonresponse
Square footage	6751	1525	22.6
Square footage category	6751	196	2.9
Year construction was completed	6751	1906	28.2
Year of construction category	6751	313	4.6
Expansion or reduction since 12/31/86	5889	106	1.8
PCs/computer terminals in building	6751	170	2.5
Able to switch main heating fuel	6102	369	6.0
Percent heated in 1992	6102	216	3.5
Percent cooled in 1992	5429	195	3.6
Commercial refrigerator/freezer equipment present	6751	136	2.0
Building owner	6751	126	1.9
Building is completely vacant	6751	133	2.0
Multibuilding facility or complex	6751	115	1.7
Principal facility activity	3165	118	3.7
Occupant status	6751	115	1.7
Number of establishments/organizations	6751	246	3.6
Space vacant for at least 3 months	6751	167	2.5
Months in use out of past 12 months	6751	253	3.7
Total weekly hours open	6526	418	6.4
Total weekly hours open category	6526	79	1.2
Heat/cool equipment in use extra hours	5649	208	3.7
Lighting equipment in use extra hours	5649	206	3.6
Number of workers (all shifts)	6526	1230	18.8
Number of workers category (all shifts)	6526	217	3.3
Number of workers	6526	1139	17.5
Number of workers category	6526	224	3.4
Wall construction material	6751	126	1.9
Roof construction material	6751	254	3.8
Building shape	6751	119	1.8
No. ext. walls attached other structure	4972	105	2.1
Percent glass on exterior	6751	280	4.1
Percent lit during operating hours	6751	288	4.3
Percent lit during off-hours	6751	366	5.4
Variable air volume (VAV) system	6751	373	5.5
Economizer cycle	6751	299	4.4
Roof or ceiling insulation	6751	427	6.3
Exterior wall insulation	6751	631	9.3
Storm windows or doors	6751	229	3.4
Tinted or reflective glass	6751	178	2.6
Shadings or awnings	6751	174	2.6
Most windows can be opened and closed	6751	161	2.4
Utility sponsored DSM, past 3 years	6751	1081	16.0
Building participated DSM, past 3 years	6751	546	8.1
Facility participated DSM, past 3 years	3110	287	9.2
Building plans participate in DSM in future	6751	873	12.9
Energy audit ever performed	6751	949	14.1
Regular preventive maintenance program	6751	226	3.3
Energy management and control system	6751	180	2.7
Other features to help conserve energy	6751	207	3.1
Person responsible for HVAC equipment	6751	152	2.3
Non-emergency generating capability	6751	137	2.0
Central physical plant on facility	3165	100	3.2
Expenditures for electric in 1992 category	6574	1572	23.9
Expenditures for natural gas in 1992 category	4160	1157	27.8
Interruptible natural gas service	4160	505	12.1
Building uses transportation gas	4160	158	3.8

Source: Energy Information Administration, Office of Energy Markets and End Use, 1992 Commercial Buildings Energy Consumption Survey.

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 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 for CBECS, 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 relative standard errors (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 B, "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.

The replication method used begins by grouping 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 members 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:

$$S_{X'} = \sqrt{S_{X'}^2} \ .$$

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

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

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.

The use of the row and column RSE factors is illustrated in Appendix A, "Detailed Tables" section.

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

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

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 is the mean over all rows i for a particular column j. The row and

column RSE factors are then computed as

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

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

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. 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. Additionally, no RSE Column factors are included for the four columns of median statistics found in Appendix A, "Detailed Tables" (Table A1).