Appendix C

Estimation

Estimation refers to the process by which national (population) estimates are made based on the responses received from a sample. Estimation for the Facility Survey had to deal with three problems, each of which will be described in a separate section of this appendix.

The first estimation problem was that, although the Facility Survey was targeted at multibuilding facilities, the basic CBECS sample was designed for inferences about individual buildings. The following section, "The Network Estimator," describes how estimates and their variances were produced. The theory of network estimation guided the overall design of the Facility Survey.

The second estimation problem was that not all sampled units responded to the Facility Survey, and not all respondents completed all items. The section, "Imputing for Missing Responses," describes the procedures used to compensate for missing responses.

The third estimation problem was that of producing estimates and variances which accounted for the effects of imputation. The inference from sample to population introduced sampling variance, due to the fact that a small part of the population was selected to represent the whole. The imputation procedures introduced imputation variance, since the responses received were used to guess how the nonrespondents would have responded. Since a large portion (over a third) of the units were completely imputed, the effects of imputation could not be ignored. The third section, "Estimation With Multiple Imputation," describes the method chosen to produce both point estimates and variances from the Facility Survey data.

The Network Estimator

The facility data can be used to derive estimates of national totals across all facilities. As described by Goldberg,[16] the estimator used is a network estimator, based on the basic CBECS sample design. For any aggregate quantity Q defined as the sum of facility quantities Q_f, the estimator is

$$\hat{Q} = \sum_{f} Q_f \sum_{b \in F_f} (s_b/S_f)\, d_b w_b \qquad \textbf{(1)}$$

where

F_f = the set of multibuilding facilities with central physical plants,

s_b = square footage of sample building b,

S_f = total square footage of in-scope buildings on facility f,

d_b = 0/1 indicator variable for whether building b is in the responding sample, and

w_b = sampling weight for building b (including unit nonresponse adjustment).

[16] Miriam L. Goldberg, "An Adjunct Facilities Survey for a Complex Buildings Survey," *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, (1989).

Nonsampled and nonresponding buildings do not contribute to the sum $\hat{Q}$, since $d_b = 0$ for these buildings. Likewise, a facility f with no responding buildings does not contribute to the sum, since the multiplier for the quantity Q_f is zero.

The estimator $\hat{Q}$ can also be expressed as

$$\hat{Q} = \sum_{f} Q_f W_f \qquad \textbf{(2)}$$

where

$$W_f = \sum_{b \in F_f} (s_b/S_t)\, d_b w_b \qquad \textbf{(3)}$$

is the derived facility sample weight for facility f.

Alternatively, the estimator can be expressed as a sum over all sample buildings:

$$\hat{Q} = \sum_{b} (Q_{f(b)} s_b/S_{f(b)})\, d_b w_b, \qquad \textbf{(4)}$$

or

$$\hat{Q} = \sum_{b} (Q_{f(b)}/S_{f(b)})\, s_b d_b w_b, \qquad \textbf{(5)}$$

where f(b) indicates the facility f to which building b belongs. For example, $Q_{f(b)}$ is the quantity for the facility to which building b belongs.

Equation (4) indicates that the facility total Q_f is allocated among sample buildings b on the facility in proportion to the buildings' floorspace. This allocation is not intended as an accurate estimate of the individual buildings' proportion of central plant outputs. Indeed, the building may receive no outputs from the central plant, in which case this proportion is known to be zero. Rather, the allocation implied by the estimator $\hat{Q}$ is a statistical apportionment that yields an unbiased estimate for the aggregate Q.

Equation (5) shows that the estimator does not require explicit values for the facility consumption Q_f and the facility eligible floorspace S_f, only for the ratio of the two. Thus, for facilities where both the consumption amounts and the eligible floorspace are missing, it is less critical to impute these items separately than to get a reasonable imputation for their ratio. For instance, it may be more stable to hot-deck the consumption per square foot than to impute consumption and floorspace separately.

Ignoring unit nonresponse adjustments[17], the building sampling weight w_b is simply the reciprocal of the sampling probability. With the inflation for unit nonresponse, we can consider the adjusted weight w_b as the reciprocal of the probability p_b that a building both is sampled and responds to the survey.

This probability can be partitioned as the product of the probability P_f that at least one building from the facility is in the (responding) sample and the conditional probability $p_{b|f}$ that the building b is sampled, given that facility f is in the sample. This partition is valid even though the two components are not separately known in general.

[17] The response rate for the 1989 CBECS was 92.5 percent, so that the effects of the nonresponse adjustments are small.

The derived sampling weight W_f (Equation (3)) estimates the reciprocal of the facility probability P_f. This approach provides an unbiased estimator of the facilities aggregate q using building sampling weights only, without explicit calculation of exact facility sampling probabilities.

Given that the facility f is in the responding sample, the conditional expectations are $E_{|f}(d_b) \; = \; p_{b|f}$, and

$$E_{|f}(W_f) \; = \; 1/\mathbf{P}_f . \tag{6}$$

Over all possible samples, $E(d_b) = p_b = 1/w_b$, so that $E(W_f) = 1$. Thus, each facility f has expected contribution Q_f to the facility aggregate estimator $\hat{Q}$, and

$$E(\hat{Q}) \; = \; \sum_f Q_f \; E(W_f) \; = \; \sum_f Q_f \; = \; Q. \tag{7}$$

Therefore, $\hat{Q}$ is unbiased.

Imputing for Missing Responses

As described in Appendix B, the Facility Survey experienced two types of nonresponse, unit and item. If the sole purpose of the Facility Survey were to improve the quality of estimates from the District Heating and Cooling Suppliers Survey, then nonresponse would be a problem only insofar as it limited the amount of data verification that could be accomplished. However, the Facility Survey was also designed to provide population estimates of inputs and outputs of central plants. If untreated, unit and item nonresponse could lead to serious biases in these estimates.

Imputation Versus Reweighting

Unit nonresponse to the Facility Survey could be handled either as unit nonresponse or as item nonresponse. In the CBECS, nonresponse to the Building Characteristics Survey is considered unit nonresponse. Sampling weights are adjusted for unit nonresponse within cells defined by sampling information. On the other hand, nonresponse by a building's energy supplier to the Energy Suppliers Survey is treated as item nonresponse. The missing items are imputed using data from the Building Characteristics Survey.

It seemed preferable to treat Facility Survey unit nonresponse as item nonresponse, as with unit nonresponse to the CBECS Energy Suppliers Surveys. From the Building Characteristics Survey, there is abundant information about the characteristics of the sampled buildings on the facility, which can be used to impute facility information. In addition, the facility activity is known for all sampled facilities, including unit nonrespondents. These activities were obtained by examining and coding the questionnaires of the sampled buildings on the facility.

There would be several problems associated with a reweighting approach to unit nonresponse adjustment. Both nonresponse adjustment and variance estimation would be difficult if Facility Survey nonresponse were treated as unit nonresponse. The usual unit nonresponse reweighting adjustment is based on sampling stratification cells. The adjustment cells used by CBECS are based on building-level characteristics. Each facility could have several sampled buildings, each one in a different cell, making the assignment of a facility to a cell problematic.

The building-based sampling stratification has similar consequences for variance estimation. CBECS variances are estimated using a jackknife procedure, with buildings assigned to replication units based on the sample stratification cells. It is conceivable that buildings associated with the same facility could be assigned to different replication units.

Finally, it would be desirable for each building that belonged to a multibuilding facility with a central plant to have the associated Facility Form information on the final building data file. Using nonresponse adjustment factors would leave a substantial fraction of applicable buildings without these data.

For all the above reasons, nonresponse to the Facility Form was treated as item nonresponse.

The Hot-deck Procedure

The two basic CBECS techniques for handling item nonresponse are regression modeling and hot-decking. Regression modeling is used for missing energy consumption. The CBECS estimates for consumption of electricity, natural gas, fuel oil, and district heat are the most important estimates provided by the entire survey. Hot-decking is used to impute for missing building characteristics, which include a large number of less critical items. Developing separate models for each item with missing values would be time consuming, without having a major impact on data quality.

About 50 Facility Survey items, some more important than others, required imputation. Some items, such as the quantities of energy input and output, might have been modeled individually. However, the process would have been time consuming, with no guarantee of success. Therefore, hot-decking, the easier of the two techniques to implement, was chosen to impute for all missing Facility Survey data.

In the hot-deck method, cases are divided into two groups: "donors" (with reported values for the items of interest) and "receivers" (with missing values for the item of interest). The separation into donors and receivers is usually done within "cells" formed by cross-tabulating items known for both groups. Within cells, values are considered to be missing at random, and imputation is performed by randomly selecting a donor and copying its value onto the receiver.

Facilities are matched on some or all of the relevant characteristics, such as facility activity, size, types of input and output energy, power generation characteristics, and the type of fuels listed as being delivered to sampled buildings on the facility. Numeric items are not hot-decked directly. Instead, hot-decking was performed on ratios between nonmissing quantities and missing quantities, such as outputs to inputs or in-scope square footage to total square footage.

In hot-deck imputation, the definition of cells constitutes an implicit form of modeling. Cell variables, analogous to main effects in the analysis of variance, are chosen for their relationship with the target item, and all interactions between cell effects are included.

Some facilities with reported data for particular items were discovered during data editing to have impossible or implausible answers for those items. Those cases were disqualified as donors for the items in question.

The CBECS uses a vector hot-deck procedure. With this procedure, the facility that donates a particular item to a receiver also donates related items (up to 5) if any of these are 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, and avoids implausible combinations of facility responses.

The Facility Survey data were organized into four data groups for imputation: (1) facility characteristics, (2) input and output fuels, (3) output amounts, and (4) input amounts. These four groups will be discussed in turn.

Facility Characteristics

The first facility characteristic to be imputed was whether the facility was a multibuilding facility. The matching variables for this critical item were (1) whether there was more than one sampled building on the facility, (2) whether the sampled building(s) reported receiving any energy from the central plant, and (3) whether the facility was industrial. The remaining items imputed in the first group were items 2 through 8 from the Facility Form, facility size and power production characteristics.

The facility size items (total and in-scope number of buildings and floorspace) were not imputed directly. Instead, the following four ratios were imputed:

- the ratio of the floorspace of sampled buildings from the Building Characteristics Survey to the in-scope floorspace from the Facility Survey;

- the ratio of the in-scope floorspace to the total facility floorspace;

- the ratio of the number of sampled buildings to the number of in-scope buildings;

- the ratio of the number of in-scope buildings to the total number of buildings on facility.

Since the number and floorspace of the sampled buildings was known (or previously imputed), the imputed ratios could be used to fill in whichever items were missing. Matching variables used to impute these ratios included whether the facility was industrial, as well as categorical versions of the ratios.

The facility power production characteristics were imputed matching on each other, and also on information from the Building Characteristics Survey on energy sources used in the sampled buildings.

Input and Output Fuels

The second group consisted of the input and output fuel use indicators (binary variables indicating whether a fuel was used or produced). Output fuel indicators were imputed first, using building data on the energy sources supplied to the buildings by central plants. Also used were a categorical version of the ratio of sampled to total buildings on the facility (an indicator of how complete the sampled buildings' energy sources might be) and whether the facility had a cogeneration system. Input fuel indicators were imputed using the output fuels, plus the Census region (to reflect regional availability of fuels).

Once imputations for input and output fuel indicators were completed, input and output amounts could be imputed.

Output Amounts

The third group comprises the facility output quantities. Output amounts were imputed before input amounts because available information on the size of the loop served by outputs could help establish the demand for the outputs. Input amounts could then be sized based on the outputs required.

As was the case for the facility size items in the first group, output amounts were not imputed directly. Instead, the following two ratios were imputed for each output energy source:

- the ratio of the output loop floorspace to the total facility floorspace;

- the ratio of the output amount to the output loop floorspace.

These ratios were imputed using the input and output fuel use indicators, as reported or imputed in the second group of items. For output amount per loop floorspace, climate zone was an additional variable. Since the value of total facility floorspace was either reported or had been imputed in the first group, these ratios could be used to obtain the output amount.

Input Amounts

Finally, the fourth group consisted of the facility input quantities. For inputs, two types of ratios were used:

- overall: the ratio of the total input Btu to the total output Btu;

- for each input fuel: the ratio of the Btu of the input fuel to the total input Btu.

Values less than 0.1 or greater than 1.5 were excluded as donors. Both input and output fuel use indicators were used to impute these ratios.

Estimation with Multiple Imputation

Because of the high Facility Form nonresponse rates, the final facility data set was heavily imputed. Final estimates were based on responses for about one-half to two-thirds of the in-scope Facility Survey cases. This limitation would exist regardless of what strategy was adopted for handling the Facility Form nonresponse. The imputation strategy used was multiple imputation.

Multiple imputation is defined as "the technique that replaces each missing or deficient value with two or more acceptable values representing a distribution of possibilities."[18] The multiple imputation method involves two or more independent replications of the imputation methodology. Each replication completes the full-sample data set by imputing once for each missing value. The analysis and variance estimation then proceeds using standard survey methods. The multiple full-sample estimates are combined to obtain the overall survey error, including the contribution due to imputation effects.

In reflecting the "distribution of possibilities," multiple imputation offers two main advantages. First, for point estimates, multiple imputation makes better use of the available data than does single imputation. Second, multiple imputation allows variances to include uncertainty due to item imputation in overall estimates of survey error. Previously, multiple imputation had been used in CBECS for the 1989 building characteristics data in a limited evaluation of the imputation procedure and its effect on variances.[19]

There are two stages in the implementation of multiple imputation: (1) the imputation phase, in which two or more independent imputations are made for each missing value, and (2) the estimation phase, in which the multiple imputations are used to estimate the quantities of interest and their variances.

[18] Donald B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, (New York, Wiley, 1987), p. 2.

[19] Eugene M. Burns, "Multiple Imputation in the 1989 Commercial Buildings Energy Consumption Survey: Building Characteristics," CBECS Technical Note 86, Energy Information Administration, Office of Energy Markets and End Use (April 8, 1991); Energy Information Administration, Office of Energy Markets and End Use, *Commercial Buildings Characteristics 1989*, DOE/EIA-0246(89) (Washington, DC, June 1991), Appendix B.

Imputation

The first step in multiple imputation is generating of multiple, independently drawn, imputed values. To ensure that imputations are independent, both within and between sets of imputations, CBECS employs a type of hot-decking known as approximate Bayesian bootstrapping.[20]

In a cell with n_d potential donors and n_r receivers, the approximate Bayesian bootstrapping first requires a random draw of size n_d, with replacement, from the set of donors. (The only restriction on cell size was that n_d must be greater than or equal to 2.) If the hot-deck cell definition constitutes a satisfactory implicit model, then the quality of the resulting imputed value is not affected by this randomization within cells. Next, n_r imputed values are randomly drawn, with replacement, from the pool of sampled donor values.

Multiple imputation was accomplished simply by repeating the procedure using different seeds for the random number generator. The result was multiple sets of imputed values, which could be used to form multiple completed versions of the full data set. In all, 10 sets of imputed values were produced, the upper range of the 2 to 10 suggested as a reasonable number by Rubin.[21]

Estimation

The second part of multiple imputation is incorporating the m sets of imputed values into the estimation process. The m sets of imputed values update the unimputed file to form m completed data sets. With multiple imputation, the survey's standard methods of producing point estimates and variances can be applied to each completed data set. The resulting sets of estimates and variances are then combined into an overall set of estimates and variances which incorporate imputation effects.

In the standard CBECS methods, a completed full-sample data set is used to obtain point estimates of totals as follows:

$$\hat{X} = \sum_{i=1}^{n} w_i x_i, \qquad \textbf{(8)}$$

where w_i is the overall full-sample adjusted weight for the i^{th} building, x_i is the value of the variable of interest for the i^{th} building (ignoring distinctions between reported and imputed values), and n is the number of buildings included in the cell. Estimates for the number of buildings are formed by summing the sampling weights, i.e., by letting x_i=1 for all n buildings.

Due to the complexity of the sample design, the CBECS uses the jackknife replication method (with 40 collapsed strata) for variance estimation. To capture variation due to unit nonresponse, weight adjustment is performed separately within each replicate, as well as overall. The 40 sets of replicate weights are used to compute mean square errors about the full-sample point estimates, as follows:

$$\hat{S}^2 = \sum_{k=1}^{40} (X_k - \hat{X})^2, \qquad \textbf{(9)}$$

where X_k is the point estimate based on the k^{th} replicate, and $\hat{X}$ is the point estimate based on the full-sample data set. The replicate totals are calculated as

$$X_k = \sum_{i=1}^{n} I_{ki} w_{ki} x_i, \qquad \textbf{(10)}$$

[20] Donald B. Rubin and Nathaniel Schenker, "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association* 81, 394 (June 1986), pp. 366-374.

[21] Donald B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, (New York, Wiley, 1987), p. 15.

where w_{ki} is the kth replicate adjusted weight for the ith building, and the replicate inclusion indicator, I_{ki}, takes the value

0, if the ith building is in the kth stratum in the unit omitted from the kth jackknife replicate,

2, if the ith building is in the kth stratum in the unit included in the kth jackknife replicate, and

1, for all buildings not belonging to the kth stratum.

The standard methods described above are applied to each of the m completed data sets. The combined overall point estimate for each item is obtained as the mean of the full-sample estimates, $\bar{X}_m$, as follows

$$\bar{X}_m = \sum_{r=1}^{m} \hat{X}_r / m, \tag{11}$$

where m is the number of completed full-sample data sets, and $\hat{X}_r$ is the full-sample point estimate for the rth completed full-sample data set (Equation (2)).

The combined overall variances are estimated as the sum of two components:

- a within-completed data set component, $\bar{W}_m$, calculated as the mean of the full-sample variances,

$$\bar{W}_m = \sum_{r=1}^{m} \hat{S}^{2}{}_{r} / m, \tag{12}$$

where $\hat{S}^{2}{}_{r}$ is the mean square error estimate (Equation 9) for the rth completed full-sample data set, and

- a between-completed data set component, B_m, estimated as the variance of the full-sample point estimates,

$$\mathbf{B}_m = \sum_{r=1}^{m} (\hat{X}_r - \bar{X}_m)^2 / (m-1). \tag{13}$$

The total variance, V_m is

$$V_m = \bar{W}_m + (1 + m^{-1}) \mathbf{B}_m. \tag{14}$$

where the factor $(1 + m^{-1})$ is an adjustment for the use of a finite number of imputations.

The standard CBECS estimation methods have been incorporated into an EIA modification of TPL, TPL/VARIANCE, which produces publication-quality tables of estimates and their associated relative standard errors.[22] Incorporating multiple imputation in estimation would require further modification of TPL/VARIANCE. The estimates presented in this report were programmed in SAS using WESVAR[23]1 to produce the standard survey estimates.

[22] Paul M. Gargiullo, *TPL-VARIANCE: System Documentation*, Energy Information Administration, Office of Energy Markets and End Use; Paul M. Gargiullo and Miriam L. Goldberg, "A Modified Table Producing Language (TPL) System for Producing Tables of Survey Statistics with Variances," *Proceedings of the Bureau of the Census Fifth Annual Research Conference*, (Washington, DC: Bureau of the Census, 1989).

[23] Paul Flyer and Leyla Mohadjer, *The Wesvar Procedure* (Rockville, MD: Westat, Inc., 1988).