AN ADJUNCT FACILITIES SURVEY FOR A COMPLEX BUILDINGS SURVEY

Miriam L. Goldberg, Energy Information Administration[1]
1000 Independence Avenue, SW (EI651) Washington, DC 20585

KEY WORDS: Complex sample, network sample, multiplicity estimator, energy

INTRODUCTION

Relating summary information about a particular population of sampling units to results for some higher level of organization is often a challenge for complex surveys. For example, it may be of interest to compare results from a survey of individuals to results from a survey of households. In the Commercial Buildings Energy Consumption Survey (CBECS), data are collected and summarized for buildings, but for some purposes it is more natural to collect and analyze data for facilities such as college campuses and hospital complexes.

In this paper, an approach is described to relating building-level statistics to estimates for facilities, as an example of a method that may be more generally useful. In the general case, a base sample of individuals is supplemented with an adjunct sample of associations. The adjunct design described here can be viewed as a special case of a network sample.

The CBECS is conducted triennially by the U.S. Energy Information Administration (EIA). The purpose of the survey is to provide estimates of energy consumption in commercial buildings, and to characterize this buildings population in terms of features related to energy consumption patterns. The survey uses a complex multi-stage probability sample, with the individual building as the sampling unit.

An adjunct survey of facilities has been designed for the next cycle of the CBECS, to expand the survey's scope without altering its basic structure. A facility is defined as a group of buildings on the same site, owned or operated by a single organization or person.

The adjunct survey has three related objectives. The first is to capture a component of commercial-sector energy use that takes place in noncommercial buildings located on commercial facilities. Of particular interest is energy consumption and output of commercial cogeneration units. These units are relatively rare, but are anticipated to be of increasing importance. The building housing a central electricity generation plant on a campus complex would typically be classified as manufacturing, and hence be excluded from the basic survey as noncommercial.

The second purpose of the adjunct survey is to provide a basis for better building-specific estimates of district heating and cooling (DHC) energy consumption, by obtaining information at the power plant level, rather than only at the building level. The CBECS collects energy consumption data for each sampled building from the building's energy suppliers. In the case of a building supplied with DHC from a central plant on the same facility, the central plant is the energy, and data, supplier. Frequently in such cases, total fuel inputs to the central plant and outputs from it are available, but no building-specific consumption.

The third purpose is to help identify the overlap between the commercial survey, where the sampling unit is a building, and the Manufacturing Energy Consumption Survey (MECS), where the sampling unit is an establishment or facility.

SCOPE OF THE BASIC AND ADJUNCT SURVEYS

The target population of the basic CBECS consists of commercial buildings. Commercial includes any activity that is neither industrial nor residential. For practical reasons, only buildings larger than 1,000 square feet are included. Commercial buildings may be located on noncommercial facilities, as in the case of the administrative offices on a manufacturing site. Conversely, a commercial facility such as a hospital complex or university campus may include noncommercial buildings such as residences or a central power plant. The basic CBECS covers commercial buildings on both commercial and noncommercial facilities, though coverage is relatively inefficient on the noncommercial facilities. The basic survey

[1]The author thanks Dwight French and Eugene Burns for helpful discussions. The opinions expressed herein are those of the author and should not be construed as representing the opinions or policy of any agency of the United States Government.

excludes noncommercial buildings, even on commercial facilities.

The adjunct survey will expand CBECS energy consumption coverage to include central plants on commercial facilities, even if these plants are housed in buildings whose principal activity is noncommercial. Ideally, it might be of interest to expand the survey to capture all commercial-facility consumption. However, such an expansion of the survey's scope is not practical within the resources available. Instead, the adjunct will pick up one large component of commercial-facility consumption currently not covered by either the CBECS or another EIA consumption survey. Consumption in noncommercial buildings, other than central plants, on commercial facilities will still not be covered by any of these surveys.

The basic CBECS population definition, sample design, and estimation procedures will be essentially unchanged by the addition of the facilities adjunct. Central plants on commercial facilities will be eligible for the basic CBECS only if the plant satisfies the definition of a commercial building.

A facility will be included in the adjunct sample if one or more buildings on the facility were included in the basic CBECS sample. As a cost-saving measure, the adjunct is further restricted to include only those facilities with central power plants. Because the number of facilities included in the adjunct sample is expected to be relatively small, facility estimates will be provided only at high levels of aggregation.

THE GENERAL ADJUNCT SURVEY

Extending the CBECS example to the more general case, an adjunct survey can be defined as follows. Let U be a population of individuals, or base sampling units, u, and let S be a sample from U. Assume that each base sampling unit belongs to a unique association A, where A is a subset of U. That is, the set U_A of associations A forms a partition of the base population U. The adjunct sample S_A of associations consists of all associations A containing at least one base unit u belonging to the base sample S. The adjunct sample may be further restricted to include only those associations A that satisfy certain criteria. That is,

$$S_A = \{A: A \cap S \neq \phi\} \cap C_A ,$$

where C_A is the criterion set of associations that satisfy the restrictions.

In terms of the CBECS, the base population U is the universe of commercial buildings u, and the association population U_A is the set of facilities. Buildings that are not part of a larger facility can be thought of as single-building facilities. The criterion set C_A is the set of multi-building facilities that have a central physical plant. Other examples of surveys that would fit this framework include an adjunct sample of households taken from a base sample of individuals, or an adjunct sample of employers taken from a base sample of members of the labor force. Another possible application related to the CBECS would be to extend the EIA's Residential Energy Consumption Survey (RECS), which is a base sample of housing units, to add an adjunct sample of buildings. The buildings adjunct for the RECS would help delineate the boundaries between that survey and the buildings-based CBECS, just as the facilities adjunct to the CBECS will help to delineate the boundaries between it and the establishment-based MECS.

The adjunct sample is very similar to a traditional network sample (Sirken, 1972 and Sudman et al., 1988). Network sampling is typically used for sampling individuals that are rare or difficult to locate, such as victims of rare diseases, or transients. To obtain this sample of target individuals, a sample of more easily identified enumeration sources is drawn. Each target individual is linked to one or more enumeration sources. The sampled enumeration sources supply information on the target individuals they know about.

In the adjunct sampling framework, the association corresponds to the target individual, and the base sampling units to the enumeration sources. For the adjunct, each base sampling unit belongs to one and only one association. This corresponds to a special case of network sampling, where each enumeration source can report on at most one target individual.

A conceptual distinction between the adjunct sample and traditional network sampling is in the focus of the sampling effort. In common network sampling applications, the sample of enumeration sources is designed to capture target individuals as efficiently as possible. Optimal design with respect to this objective is an important consideration of network sampling theory. For the adjunct survey, the base sample is assumed to

be pre-existing, and to be designed primarily to cover the base population as efficiently as possible. The adjunct sample is taken as a low-cost supplement to this pre-existing base survey.

DESIGNING THE ADJUNCT SAMPLE

One approach to designing the adjunct sample would be to treat it as separate from the base sample, except in the identification of members of the adjunct sample. With this approach, sampling weights would be developed for each sampled association, after which estimates for the association population would be computed with no reference to the base sample. The design in this case would have to include provision of the supplementary information needed to determine association sampling weights.

An alternate approach is to use the types of multiplicity estimators common to network sampling. This approach bypasses the need to develop explicit weights for the adjunct sample, but requires other supplementary information to compute the multiplicity estimators. The two approaches are illustrated with reference to the CBECS.

The basic CBECS sample consists of an area sample supplemented by a list sample. The area sample alone can provide comprehensive, unbiased estimates for the target population. However, estimates based only on the area sample would have a high variance because of the highly skewed buildings population. Buildings with very high consumption occur rarely in the population, hence are included rarely in the area sample, but account for a large proportion of energy consumption. For this reason, a supplementary sample is drawn from special lists, assembled for each Primary Sampling Unit (PSU), of potentially high-consuming buildings.

Unbiased estimates are constructed from the combined area and list sample by setting within-PSU selection probabilities to 1 for buildings in the overlap between the list frame and the area sample. That is, any building selected for the area sample that was on the special lists, even if it was not sampled from the lists, is treated as a conditional certainty building. This procedure not only gives unbiased estimates of population aggregates (Chu, 1988) but also yields appropriate variance estimates (Gargiullo and Goldberg, 1988).

This overall structure to the basic CBECS design affects the approach to designing the adjunct survey.

DESIGN APPROACH I: DEVELOPING ASSOCIATION WEIGHTS

A conceptually simple approach to developing the facility adjunct is to assign a sampling weight, equal to the inverse of the selection probability, for each facility identified through the basic CBECS. Nationwide facility aggregates would then be estimated by multiplying facility attributes by facility weights, and summing across the adjunct sample. The major question to be worked out for this approach would be how to determine the overall selection probability for each facility identified. At a minimum, greater care would be required at the listing stages, for both the area and list frames, to identify facilities before sampling.

In principle, if a facility has been identified prior to sampling, the facility selection probability can be determined from the building sampling information. In many cases, a listing selected from the special list corresponds not to a single building, but to a facility, which is then subsampled for the CBECS. In this case, the selection probability for the facility itself is known. If the individual buildings in a facility are each listed separately, but all are known and listed, the overall probability of selecting at least one is in principle computable, though not necessarily easily.

More problematic are facilities that are identified as such only at the time of interview. No matter how carefully the listing is done, some cases like this are bound to arise. Facilities identified only in the field from list-sampled buildings could simply be excluded from the facilities sample, with some loss of efficiency but no bias.

For facilities identified in the field from area-sampled buildings, the interviewer would have to collect some additional information to allow selection probabilities to be determined. This determination and the associated information required could become quite complicated if the facility spanned more than one segment (third-stage area sampling unit).

This problem could be mitigated somewhat by assigning each facility to a single segment. In the case of a facility spanning two segments, the assigned segment would be the one containing most of the facility's buildings. In the case of a tie, or in the case of a facility spanning more than two segments, the segment for the facility would be randomly assigned with equal probability from the segments spanned. The interviewer's

additional tasks would then be: (1) to identify on the sampled building's segment list all other buildings on this facility; (2) to ascertain the total number of buildings on the facility; and (3) to identify which other segments or unsegmented SSU's the facility spans.

On the basis of this information, the facility would be assigned either to a segment containing one of its area-sampled buildings, or to another segment, or unsampled SSU. The facility would be included in the adjunct sample only if it was assigned to a segment from which one of its buildings was sampled. In such a case, the identification of facility buildings on the segment lists would allow computation of the overall inclusion probability for the facility.

While this procedure seems fairly complicated, in practice it should usually involve facilities with only a few buildings, since larger facilities are likely to show up on the special lists. Facilities picked up from the area sample that are also on the list frame could be handled by procedures similar to the overlap adjustment procedure for the basic CBECS. It is likely, though, that a relatively large proportion of the area-based facilities sample would end up in the overlap sample, for reasons just noted. The implications for variance and variance estimation of a large overlap sample would need to be considered.

DESIGN APPROACH II: MULTIPLICITY ESTIMATORS

To avoid many of the logistical difficulties described above, an alternate approach has been adopted. This approach is to develop facilities estimators that rely only on basic CBECS building weights, without requiring explicit facilities sampling weights. This alternate approach does introduce some other logistical problems, but these should be more manageable.

For the general case, let T_k be the value of a particular attribute for association A_k. For the CBECS, T_k might be total central-plant natural gas consumption for the facility, or a dummy variable for a particular facility activity type. Further, let

N_k = the number of base units u in association A_k

p_{kj} = the probability that base unit j from association A_k
is included in the base sample S

d_{kj} = 0/1 indicator for the inclusion of base unit j from association A_k in the base sample.

The sum T of attribute T_k over all associations can be represented as

$$T = \sum_{k} \sum_{j=1}^{N_k} T_k/N_k \qquad (1)$$

Corresponding to expression (1) we can construct a number-based estimator for the aggregate

$$T\# = \sum_{k} \sum_{j=1}^{N_k} (T_k/N_k)d_{kj}/p_{kj}. \qquad (2)$$

That is, for each sampled base unit in an association, the association attribute divided by the number of units in the association (the "attribute per unit") is weighted by the building weight, $1/p_{kj}$. In terms of the CBECS, for each sampled building in a facility, the attribute per building is weighted by the building weight.

The number-based estimator T# is unbiased, provided that T_k and N_k are reported without error, and all buildings included in the count N_k have positive probabilities of entering the sample.

For the CBECS, one problem that arises with this approach is that typically some buildings included in the reported building count for a facility will be noncommercial, hence have zero selection probability. Thus, for example, an agricultural building on a college campus would be included in the denominator N_k in T#, but would never be represented in the sum of weights p_{kj} over sample buildings. Overall, the effect of including buildings in the facility total that are excluded from the buildings sample is to bias the facilities aggregate downward.

There are essentially two ways to address this potential bias problem, namely by changing the buildings eligible for the sample to match the facility total, or by changing the definition of the facility total to match the buildings eligible for the sample. The first approach would mean retaining in the facilities estimator any building sampled on a commercial facility, even if the building itself turned out to be noncommercial or out of scope based on size. In addition, the listing process would have to include all buildings on commercial facilities, even if clearly out of scope for the basic CBECS, to make these buildings available for the facilities estimator. Such buildings would still be excluded from the basic CBECS estimator. However, the interviewer

would need to collect some information from each such building, and the facility questionnaire would be administered at a site, even if the only sampled buildings on the site was CBECS-ineligible. This collection of limited information on ineligible buildings would entail extra interviewer effort and respondent burden.

The approach adopted instead requires the facilities questionnaire respondent to report building counts and floorspace specifically for CBECS-eligible buildings. To focus responses more effectively, the facilities questionnaire first asks for the total number of buildings, including ineligibles, then asks again, restricting to nonresidential, nonindustrial buildings over 1,000 square feet. In addition to reducing the response error, asking for both the unrestricted and restricted facilities sizes will make it possible to present statistics based on the unrestricted size, which may be a more natural measure.

SIZE-WEIGHTED MULTIPLICITY ESTIMATOR

For the CBECS, the sampling probabilities p_{kj} range from 1/1 to less than 1/3000. Because of this wide range, even within a facility, the contribution of a given facility A_k to the estimator T# has a high variance, even conditional on the inclusion of the facility in the adjunct sample. Within sampling strata defined by location and general building activity, the sampling probabilities are in proportion to size, where the size is a rough estimate of floorspace. This proportionality offers the basis for a lower variance association estimator.

An alternate expression for the association aggregate (1) is a size-weighted sum

$$T = \sum_{k} \sum_{j=1}^{N_k} (T_k/M_k)m_{kj}, \qquad (3)$$

where

m_{kj} = the value of some measure of size m for base unit j in association A_k

$$M_k = \sum_{j=1}^{N_k} m_{jk}.$$

Similarly, corresponding to the count-based estimator given by expression (2) is the size-weighted estimator T*. This estimator utilizes the attribute per unit measure of size for each sampled base unit, weighted by the product of the sampling weight and the base unit's measure of size:

$$T^* = \sum_{k} \sum_{j=1}^{N_k} (T_k/M_k)d_{kj}m_{kj}/p_{kj}. \qquad (4)$$

The size-weighted estimator should have lower variance than the count-based estimator T#, because high sampling weights $1/p_{kj}$ are balanced by low measures of size m_{kj}.

For the CBECS, the measure of size to be used in the estimator is the building floorspace. An additional advantage of the floorspace-based estimator is that it is likely to be less subject to reporting bias. The count N_k of buildings on a facility may be easily misreported if there are a number of small buildings. The total floorspace M_k, on the other hand, is not very sensitive to the inclusion or exclusion of several small buildings, hence is likely to be more accurately reported. Misreporting of the information required to determine the multiplicity can cause serious bias for multiplicity estimators in general (Czaja et al., 1986).

VARIANCE ESTIMATION FOR THE MULTIPLICITY ESTIMATORS

Variances for the basic CBECS are estimated by half-sample methods (NCHS, 1966 and 1969). Variances for the estimators T# and T* can be computed by the same methods, for the CBECS and other complex surveys.

One way to implement the half-sample methods for the adjunct design would be to assign each facility (association) drawn into the adjunct sample to a stratum and pair member, then include or exclude the entire facility from the pseudo-replicate estimators T# or T*, according to whether the facility pair member is included in the pseudo-replicate. The difficulty with this approach is that a facility may cross the boundaries of segments, which define pair members for the basic sample. Rules could be developed for assigning facilities to segments, as in Design Approach I. Developing and implementing such rules might be complicated. Moreover, the facilities pairs developed in this way would be more appropriate for estimators based on that design approach, rather than reflecting the design actually adopted for the adjunct survey.

A simpler way to implement the half-sampling for the adjunct survey is computationally

consistent with the multiplicity estimators T# and T*, and also more appropriately reflects the adjunct sample design. With this approach, the multiplicity estimators are considered as special cases of aggregate estimators from the basic sample. That is, each unit in the base sample is assigned the attribute-per unit T_k/N_k for its association, and T# is regarded as a weighted sum over all base units u, rather than as a sum over associations.

$$T\# = \sum_u (T_{k(u)}/N_{k(u)})\, d_u/p_u,$$

where k(u) is the index of the association containing base unit u, and $T_{k(u)}$ is defined as zero if that association is ineligible for the adjunct sample S_A.

Similarly, the size-weighted estimator is regarded as the weighted sum of size-apportioned attributes $T_k\, m_{kj}/M_k$:

$$T^* = \sum_u T_{k(u)}(m_u/M_{k(u)})d_u/p_u.$$

Half-sample variance estimates are then computed for the adjunct estimators T# and T* just as for any other aggregate calculated from the basic survey. Thus, in terms of analysis as well as design, the adjunct survey introduces relatively little additional complications to the basic survey. The same programs used for the base survey computations can be used for the adjunct sample, for both point estimates and variances.

ADVANTAGES OF THE ADJUNCT SURVEY

The approach outlined above to designing an adjunct survey has many attractive features. First, with the base survey in place, the association sample can be obtained at relatively low cost. Analysis of this sample can also be performed at relatively low cost, using the same programs developed for the base sample. Further, association-level information can be easily related to base-level information, since the base and association sampling units are linked from the outset. Finally, the adjunct survey can provide a basis for estimating the overlaps and gaps between surveys conducted at different levels.

REFERENCES

Chu, A. 1987. "Proof That the Assignment of Conditional Weights Will Produce Unbiased Estimates" in "Weighting Procedures for NBECS III." Technical Memorandum, Westat, Inc. Rockville, MD.

Czaja, R.F., Snowden, C.B., and Casady, R.J. 1986. "Reporting Bias and Sampling Errors in a Survey of a Rare Population Using Multiplicity Counting Rules." *Journal of the American Statistical Association 81*:411-419.

Energy Information Administration. 1989. Nonresidential Buildings Energy Consumption Survey: *Commercial Buildings Energy Consumption and Expenditures, 1986.* DOE/EIA-0318(86).

Goldberg, M.L. and Gargiullo, P.M. 1989. "Variance Estimation Using Pseudostrata for a List-Supplemented Area Probability Sample." *Proceedings of the Section on Survey Research Methods, New Orleans, LA, 1988.* American Statistical Association.

National Center for Health Statistics. 1966. "Replication: An Approach to the Analysis of Data from Complex Surveys. Vital and Health Statistics. Public Health Service Publication No. 79-1269, Series 2/No.14. Washington, DC.

National Center for Health Statistics. 1969. "Pseodoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique." Vital and Health Statistics. Public Health Service Publication No. 73-1270, Series 2/No.31. Washington, DC.

Sirken, M.G. 1972. "Stratified Sample Surveys with Multiplicity." *Journal of the American Statistical Association 67*:224-227.

Sudman, S., Sirken, M.G., and Cowan, C.D. "Sampling Rare and Elusive Populations." *Science 240:* 991-996.