

Post-Stratification Methodology for the 2002 Manufactures Energy Consumption Survey

Richard Hough and Stacey Cole
Census Bureau

I. Scope and Frame

The primary goal of the Manufacturing Energy Consumption Survey (MECS) is to provide a comprehensive set of energy consumption estimates for the manufacturing sector of the US economy. The MECS includes approximately 12,000 - 21,000 (varies by survey year) manufacturing establishments selected from the manufacturing portion of the Economic Census. The Census Bureau's Business Register (BR) is the universe of establishments for the Economic Census. For the 2002 Economic Census, the BR contained approximately 350,000 active manufacturing establishments. For conducting the Manufacturing portion of the 2002 Economic Census, the Business Register partitions these establishments into "nonmail" and "mail" groups for processing purposes. In general, single-location companies with less than five employees are categorized as "nonmail". For these establishments, no questionnaires are mailed. The Census Bureau relies upon information obtained from other Federal agencies for inclusion in the Economic Census statistics. For the 2002 Economic Census, approximately 150,000 manufacturing establishments were categorized as nonmail; collectively, they accounted for about 3% of total manufacturing output. The remaining 200,000 manufacturing establishments were mailed a questionnaire to obtain operational data for the Economic Census. These establishments comprise the mail file for the manufacturing portion of the 2002 Economic Census. For the 2002 MECS, the sample frame was defined as this mail file.

II. 2002 MECS Sample Design

Within the manufacturing sector, energy consumption exhibits a considerable degree of industry and geographic concentration. Consequently, MECS utilizes a stratified-probability-proportionate-to-size (PPS) sample design. Using the classification structure defined by the North American Industry Classification System (NAICS), there is interest in producing estimates for 37 specific industries at both the U.S. and regional levels. These specific industries are high energy consumption industries; collectively, they account for approximately 47% of the energy consumed within the manufacturing sector. Estimates for 30 higher level industry aggregations are also generated at both the U.S. and regional levels. When estimates for these two industry groupings are combined, estimates for the complete manufacturing sector can be produced. There are a total of 267 industry by region strata.

For the survey, the establishment is both the sample unit and the collection unit. In assembling the sample frame, each establishment in the frame is assigned to a single stratum based on its industry classification and State code. Each establishment is assigned a "measure-of-size" (MOS) based on its expected "cost of energy" for reference year 2001. For establishments included in the 2001 Annual Survey of Manufactures (ASM), this "cost of energy" was obtained directly from the ASM database. Approximately a quarter of the establishments in the frame have their MOS obtained directly from the 2001 ASM. For establishments that are not part of the ASM sample, "cost of energy" data from the manufacturing portion of the 1997 Economic

Census was obtained when available and adjusted to a 2001-basis. For establishments that could not be located in either the 2001 ASM or 1997 Economic Census, “cost of energy” was estimated using annual payroll and industry-specific parameters.

Within each stratum, a PPS approach is used to assign establishment level probabilities. The probability of selection for a given establishment is a function of its relative MOS within the industry by region stratum and the reliability constraint specified for the stratum.

The maximum sample size was set to 15,000 establishments. Since independent samples are selected from each stratum, by adjusting the individual stratum reliability constraints, we are able to satisfy the maximum sample size constraint. In general, the strata with large expected energy consumption were assigned tighter reliability constraints. The use of reliability constraints is a means of allocating the sample efficiently over the set of strata and should not be interpreted as a predictor of the accuracy of survey results.

The actual sample selection operation involves a “fixed” sample procedure. For each stratum, the expected sample size is derived by summing the establishment probabilities and rounding to the next larger integer. An independent sample is selected within each stratum and the resulting sample is equal to this derived integer.

III. Key Concerns Regarding the Sample Frame

There are three primary concerns with the sample frame used for MECS, (1) does the frame include all in-scope establishments, (2) are the industry classifications sufficiently accurate for stratification purposes, and (3) does the weighted MOS of the MECS sample reasonably represent the target population?

Scope

The BR classification information is used to determine the set of establishments in-scope of the MECS. While a limited number of classification updates are carried to the BR on a continuous basis, the last comprehensive classification update used the results of the 1997 Economic Census. Most of the classifications in the frame are five years old. Some portion of the establishments classified as non-manufacturers in the BR are truly manufacturers and should be included in MECS sample frame; and some portion of the establishments classified as manufacturers in the BR are truly non-manufacturers and should not be included in the MECS sample frame. Based on historical data, approximately 5-6% of the establishments mailed in the Economic Census are ultimately classified in a sector that differs from their original mailed sector. The issue of significant under-coverage of the target population is of concern. [In general, over-coverage of the frame is less of a concern because ineligible units can safely be removed from the survey estimates.]

Classification

In addition to concerns regarding properly identifying the correct set of establishments for inclusion in the frame, there are concerns regarding the assignment of the in-scope establishments to the correct industry by region stratum. Again, this issue is the result of the classification updating cycle of the BR. Many of the in-scope establishments may still be

classified as manufacturers although their primary activity within manufacturing may have changed. These shifts within manufacturing may result in an establishment being classified for sampling purposes in the incorrect industry by region stratum. When re-classified in the correct stratum for estimation purposes, the resulting variances can be adversely impacted.

Measure-of-size

The “cost of energy” data used to assign the MOS, and ultimately the establishment probabilities, are also of concern. Most of the total MOS of the complete frame is attributed to establishments whose MOS was obtained directly from the 2001 ASM (approximately 83%). However, given the volatile nature of energy prices, the correlation between a 2001-based “cost of energy” and the actual “cost of energy” incurred in 2002 may not be as strong as one would like. The correlation between the assigned MOS and the 2002 “cost of energy” for the remaining establishments in the frame is likely to be even lower because it is based on 1997 data or estimated via a model. The use of a less than optimal MOS adds a degree of variability in the survey results that cannot be directly addressed during the sample selection operation.

Collectively, these three issues have a direct impact on the quality and utility of the survey results. The post-stratification methodology described in the next section is intended to rectify any material deficiencies in the sample frame by taking advantage of auxiliary information obtained from the 2002 Economic Census.

IV. Post-Stratification Methodology

The availability of the results from the 2002 Economic Census offers us the opportunity to address many of our concerns with the representativeness of the sample frame. We have updated classification information for all manufacturing and non-manufacturing establishments included in the Economic Census. In addition, we have reported “cost of energy” data for all manufacturing establishments.

This allows us to do the following:

- § Re-define the set of establishments in the target population. Using the completely updated Economic Census database, we can re-assemble the MECS target population for each MECS industry by region stratum.
- § Develop a revised “cost of energy” control total for each industry by region sample strata using the updated target population and the reported “cost of energy” data for 2002. This population control total can then be used to adjust the MECS sample weights.
- § Identify establishments that are part of the target population that were not included in the sample frame. A coverage adjustment can be derived and applied to the MECS sample to improve the representativeness of the estimates.

Classification updating

Each manufacturing establishment in the “mailed” portion of the 2002 Economic Census is sent an industry-specific questionnaire determined by the establishment’s industry classification at the time of mailout. This industry classification could be several years old. The questionnaire requests general operational data such as employment and payroll; as well as specific products-produced information. Based largely on the response to the products-produced section, the

industry classification for the establishment is re-derived.

The industry classification for each of the manufacturing establishments included in the MECS sample was extracted from the 2002 Economic Census database and carried to the MECS database. Of the approximate 12,000 respondents in the MECS database, this classification update resulted in 956 establishments changing sample strata.

There are a small number of situations where the industry classification in the Economic Census database differed from the corresponding classification in MECS database and the Economic Census classification was not accepted. In these cases, the MECS analysts independently confirmed that the MECS industry classification was correct. The Economic Census is being reviewed by a different set of analysts with differing goals and objectives. Consequently, minor discrepancies are to be expected.

Deriving “cost of energy” control totals for the target population

Eligibility

The next step is to develop “cost of energy” control totals for the target manufacturing population using information from the manufacturing portion of the 2002 Economic Census. The Economic Census database contains a significant number of establishments that are not defined as being included in the target population of the original MECS sample. Therefore, these ineligible establishments need to be removed from our data-set prior to the derivation of the stratum-level population control totals. The following establishments were removed from the Economic Census database prior to deriving the control totals.

- § Manufacturing establishments in 2002 Economic Census that were not mailed a questionnaire: These “non-mails” are defined as being out-of-scope of the MECS.
- § Establishments that were mailed a non-manufacturing questionnaire in the 2002 Economic Census and were ultimately re-classified as manufacturers based upon their response data (sector transfers): Conceptually, the MECS sample frame does not represent this population; therefore, the sample does not, either. At this point in the process, we are adjusting the sample weights of the MECS sample to account for frame errors, not frame omissions.
- § Newly formed businesses (births) that were added to the manufacturing population subsequent to the creation of the original 2002 Economic Census mail-file: Again, the MECS sample frame does not represent these establishments.

Validation of “cost of energy”

Having defined the set of establishments that were eligible to be included in the control totals, the micro-level “cost of energy” data being used to derive the “cost of energy” control totals were reviewed. We utilized an existing ratio analysis program to identify establishments with unusual “cost of energy” to “cost of materials” ratios. Each industry was examined separately.

Approximately 50 manufacturing establishments were identified with obvious errors in “cost of energy”. For the majority of these, the correct value for “cost of energy” could be determined by simply reviewing the digital image of the questionnaire. Collectively, at the all-manufacturing level, these corrections reduced the “cost of energy” control total by approximately 0.9 percent.

In addition, there were several thousand small manufacturing establishments that had zero “cost of energy” in the Economic Census. Using the median value of the “cost of energy to cost of materials” ratio at the industry level, an estimate of the “cost of energy” was estimated for these establishments. The impact of this imputation was to raise the “cost of energy” control totals by approximately 1.0 percent.

Adjusting to the “cost of energy” control totals

For each of the 267 industry by region strata, a “cost of energy” control total is derived by simply summing the “cost of energy” value across the set of eligible establishments in the strata. At the stratum level, this represents the total “cost of energy” of the target population for MECS. In general, the estimate of a stratum control total derived by summing the weighted “cost of energy” data across the set of MECS establishments in the stratum does not equal the control total. In some cells the control total is greater than the MECS estimate; and in others, the control total is less than the MECS estimate. The objective of the adjustment procedure is to eliminate this discrepancy by modifying the MECS sample weights.

In examining options for adjusting the sample weights, we defined two constraints. First, establishments with sample weights of 1.00 should not be impacted by the methodology. These units were self-representing during the sample selection operation and should remain self-representing in the estimator. Second, in the scenario where the sample weights are being lowered because the MECS estimate is greater than the population control total, the sample weights of non-certainty establishments (sample weights > 1.00) are not allowed to drop below 1.00. In the extreme case where the MECS estimate is vastly greater than the population total, these non-certainties should at least contribute their observed response data to the estimate.

The following summarizes the process that was developed to adjust the weights
 § at each ind by region stratum, a “K” is derived as follows:

$$K = \frac{\sum_{i=1}^N x_i - \sum_{i=1}^n x_i}{\sum_{i=1}^n (w_i - 1)x_i}$$

where

N	=	stratum population size from Census
n	=	sample size in stratum after post-stratification
x_i	=	cost of energy of unit i
w_i	=	current weight of unit i

For each establishment in a stratum, the adjusted sample weight is derived as follows:

$$w_{ADJ,i} = 1 + K (w_i - 1)$$

Properties of the adjustment process

- § Establishments with an unadjusted weight of 1.00 receive an adjusted weight that is also equal to 1.00.
- § The adjusted weight of an establishment is a function of the amount of the difference between the population control total and the corresponding MECS estimate, and the unadjusted sample weight of the establishment. The larger the discrepancy between the population control total and the MECS estimate, the larger the individual adjustments to the weights.
- § Within a specific stratum, establishments with low unadjusted weights are adjusted proportionally less than establishments with large unadjusted weights. Within any stratum, the adjusted weights are a linear function of the unadjusted weights, but the graph of that function does not pass through the origin; unless the MECS estimate is equal to the population total.
- § For a given level of required adjustment, the impact of the adjustment to non-certainties is a direct function of the relative importance of the certainties (weight = 1.00) within the stratum. The more dominant the certainties, the bigger the impact to the weights of the non-certainties.

In the rare event that the denominator of K equals zero, i.e. only certainty sampling units were selected, then the adjusted weight should be as follows:

$$w_{ADJ,i} = \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^n x_i}$$

This will equal 1.00 if the stratum population consists entirely of certainties. If the population has some non-certainty non-selects and the sample only has certainties within a stratum, then the only way for the weighted sum of sampled units to equal the population total would be to adjust the weights of the certainties. If the population is skewed, i.e. certainty sampling units have dominating large values, then this adjusted weight should be close to 1.00 under that scenario.

Coverage adjustment

As we have indicated, the original MECS sample frame was assembled using the manufacturing portion of the 2002 Economic Census at the time of mailout and does not completely represent the target population. Specifically, incoming sector transfers and births are not included in the sample frame, but are included in the target population. In previous cycles of MECS, an adjustment to the sample weights was made at the stratum level to account for this slight under-coverage.

By examining various establishment flags in the manufacturing portion of the Economic Census database, we are able to identify the complete set of sector transfers into manufacturing and manufacturing births. Using the reported data for “cost of energy” for these two groups of establishments, we are able to measure the combined under-coverage and derive an adjustment factor for each stratum. There are approximately 12,000 sector transfers into manufacturing and

3,000 manufacturing births represented via this adjustment. Collectively, the coverage adjustment amounted to approximately 1.1 percent at the U.S. level.

V. Concluding remarks

The frame issues discussed in this paper are not new issues for MECS. To some extent, previous cycles of MECS have had similar concerns. However, the reference year for MECS has never coincided with the Economic Census. Therefore, any observed frame problems were limited to the MECS sample and not the entire population. This severely limited our options for corrective intervention. Establishment classifications were essentially frozen when the frame was assembled. MECS establishments determined to be non-manufacturing (exits) during the data review phase continued to contribute to the MECS estimates because there were no corresponding new-to-manufacturing establishments. Establishment reclassifications within the manufacturing sector, while unbiased, were not allowed because the MECS questionnaire did not request sufficient information to reclassify an establishment.

As stated above, this was the first time that the MECS was conducted in the same year as the Economic Census. We have attempted to maximize the benefits of conducting the two surveys during the same year. The post stratification of the sample was a major part of this effort.

VI. Questions for the Advisory Panel

1. Is the methodology consistent with established practices?
2. Are there any material methodological issues that we may have overlooked?
3. Can you suggest alternative methods that may improve the quality of the results?
4. What metrics regarding the overall impact on the estimates should be developed and provided to EIA and data users?