Appendix C

Quality of the Data

Appendix C

Quality of the Data

Introduction

All data collection activities and the population estimates produced from them are subject to a variety of errors. These errors may be broadly classified under two general types, sampling and nonsampling errors.

Sampling errors are defined as the variability in a survey estimator that arises because data used to estimate population values are collected from a sample of units rather than completely enumerating the entire population. Each possible sample produces different estimates of population values, depending on the set of respondents that are selected. Consider, for example, a sample of two units from a population comprised of three units. In this example, there exists three possible sample sets of respondents, each of which produces different estimate of the population total. The difference between the estimate calculated from one of the samples and the population total is referred to as the sampling error. Nonsampling errors, on the other hand, occur in any data collection activity, whether a sample survey or a complete enumeration of the population. Nonsampling errors may be associated with any part of a survey process except sampling and can include both random and systematic (biasing) errors. Commonly recognized sources of nonsampling error include undercoverage, random and systematic response errors, unit and item nonresponse, data processing errors, and tabulation errors. This appendix describes the effect of both sampling and nonsampling errors on data from the MECS. In addition, the measure for sampling errors of the population estimates are given. More details are presented in the methodological report for the MECS.[41]

Sampling Error

The estimated values appearing in this report were developed from a sample of the universe of manufacturing establishments and, as a result, will differ from true population values that would be obtained from a complete enumeration of the manufacturing universe. This is because the MECS sample is only one of a very large number of samples that could have been selected under the same sampling specifications. Each possible sample would yield its own unique estimates of the true population values, with the differences attributable to the particular set of establishments selected into each sample.

One measure of variability due to sampling is the square root of the average of the squared differences between the estimates that would be produced by all possible samples and the mean value of those estimates. This type of measure is commonly known as sampling error. Estimates of the magnitude of these sampling errors based on data from a single sample are provided by a statistic known as the standard error of an estimate. Standard errors for MECS estimates are directly computed from the reported data using the formula:

$$S_{\hat{Y}} = \sqrt{\sum_{i=1}^{n} y_i^2 (W_i)(W_i - 1)}\,, \qquad \textbf{(4)}$$

where $\hat{Y} = \Sigma(y_i W_i)$ is the MECS survey estimator, y_i is the reported value of characteristic Y for the ith MECS sample establishment, W_i is the final adjusted weight used to inflate the sample data to population estimates, and n is the number of MECS respondents. Justification for this formula is found in the MECS methodological report.

Estimates of standard errors have been computed from the MECS sample data for the estimated aggregate values and selected ratios appearing in this report. In the 1985 and 1988 MECS reports, measures of precision were presented separately in the form of relative standard errors (RSE), that is, the standard error divided by the estimated value to which it refers. In this report, computed RSE's for Table 1 through Table 52 are approximated in a two-factor model and are imbedded into each table as "row and column factors."

[41]*Energy Information Administration, Manufacturing Energy Consumption Survey: Methodological Report, 1985*. Although this report describes data quality in the 1985 MECS, much of the discussion still holds for the 1991 MECS.

Sampling Error from Generalized Variances

The RSE's computed using standard errors from Equation (**3**) may be efficiently modeled by a Generalized Variance procedure, which has been successfully used in several complex sample surveys conducted by EIA. This procedure provides a comprehensive means of reporting generalized relative standard errors, which minimizes the publishing space required to present standard errors, and eases reader's use of precision measures. Actual RSE's (by Equation **3**) are used for statistical tests and confidence intervals presented in the text, and for determining if a population estimate is too imprecise to publish (RSE greater than 50 percent).

The estimator used to approximate RSE's is based on a two-factor model. This model-based estimator is given as

$$\hat{RSE}_{i,j} = R_i \bullet C_j , \tag{5}$$

where R_i is the row factor for the ith row and C_j is the column factor for the jth column used to compute the generalized RSE of the sample estimate at the intersection of the ith row and j^{th} column. Since RSE's calculated by this Generalized Variance technique are approximates, confidence intervals and statistical tests of significance must also be regarded as only approximate. See Figure C1 for a specific example of computing an approximate RSE.

Derivation of Row and Column Factors

Row and column factors are derived by an analysis of variance procedure with the table of RSE's. Although analysis of variance is used to derive row and column effects from which row and column factors are computed, this Generalized Variance procedure can not be considered an analysis of variance because the primary concern here is to determine model fit rather than to analyze the effects of row and column variables on the RSE's. The two-way model is fit separately for each log transformed RSE table and is consistent for every table in this report. Because of this consistency over all tables, the model can be written in general format as

$$\log(RSE_{i,j}) = m + r_i + c_j + e_{i,j} , \tag{6}$$

where m is the grand mean of $log(RSE_{i,j})$ of a "balanced" table composed of I non-zero rows and J non-zero columns, r_i is the effect of the ith row, c_j is the effect of the jth row, and $e_{i,j}$ is the error term. Model parameters are fit by the standard formulas for Ordinary Least Squares given by Cochran and Cox.[42] For a given table of log(RSE) estimates, point estimators of model parameters are given as

$$\begin{aligned}
\hat{m} &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\log(RSE_{i,j})}{I \bullet J} = \overline{\log(RSE_{\bullet,\bullet})} \\
\hat{r}_i &= \frac{\sum_{j=1}^{J}\log(RSE_{i,j})}{J} - \hat{m} = \overline{\log(RSE_{i,\bullet})} - \hat{m} \\
\hat{c}_j &= \frac{\sum_{i=1}^{I}\log(RSE_{i,j})}{I} - \hat{m} = \overline{\log(RSE_{\bullet,j})} - \hat{m} .
\end{aligned} \tag{7}$$

[42]Cochran, William G., and Cox, Gertrude M. (1957), *Experimental Design* (2nd ed.), New York: John Wiley & Sons, Inc.

Figure C1. Calculation of Generalized Relative Standard Error (RSE)

Table 1. Total Primary Consumption, 1991

SIC Code	Industry Groups and Industry	Total (trillion Btu)	Net Electricity (million kWh)	Residual Fuel Oil (1000 bbls)	Distillate Fuel Oil (1000 bbls)	Natural Gas (billion cu ft)	LPG (1000 bbls)	Coal (1000 short tons)	Coke and Breeze (1000 short tons)	Other (trillion Btu)	**RSE Row Factor**
	RSE Column Factors:	**0.6**	**0.6**	**1.3**	**1.3**	**0.7**	**1.1**	**1.2**	**1.5**	**1.1**	
		Total United States									
20	Food and Kindred Products	956	49,536	4,317	2,968	W	1,433	6,913	W	W	**7.2**
2011	*Meat Packing Plants*	49	3,410	170	252	31	157	27	0	2	**9.9**
2033	*Canned Fruits and Vegetables*	44	1,375	290	131	35	126	Q	0	*	**10.4**
2037	*Frozen Fruits and Vegetables*	40	3,071	321	76	25	41	0	0	1	**14.9**
2046	*Wet Corn Milling*	140	4,054	29	31	51	1	3,051	W	W	**11.8**
2051	*Bread, Cake, and Related Products*	32	2,240	*	131	23	23	0	0	*	**12.4**
2063	*Beet Sugar*	67	386	W	30	18	5	1,901	W	*	**5.4**
2075	*Soybean Oil Mills*	51	1,616	42	31	25	5	592	0	7	**3.5**
2082	*Malt Beverages*	50	2,328	419	58	22	8	706	0	1	**10.8**
21	Tobacco Products	24	1,002	135	40	4	23	692	0	*	**6.6**
22	Textile Mill Products	274	29,532	1,966	1,064	105	629	1,362	0	13	**7.0**
23	Apparel and Other Textile Products	44	5,645	Q	142	18	159	88	0	1	**17.9**
24	Lumber and Wood Products	451	17,878	333	2,753	39	1,009	92	0	325	**14.3**
25	Furniture and Fixtures	68	4,915	184	163	18	255	157	0	26	**20.1**
26	Paper and Allied Products	2,506	58,896	24,883	1,593	W	1,379	13,252	W	W	**4.2**
2611	*Pulp Mills*	300	2,537	4,500	162	32	141	331	0	221	**14.2**
2621	*Paper Mills*	1,211	32,735	13,455	W	252	616	8,634	W	555	**3.0**
2631	*Paperboard Mills*	859	10,396	W	W	W	W	W	0	505	**4.5**
27	Printing and Publishing	108	15,629	50	318	47	181	0	0	4	**12.6**
28	Chemicals and Allied Products	5,486	129,093	W	2,412	2,163	W	W	423	1,019	**6.1** ◄
2812	*Alkalies and Chlorine*	160	10,718	W	43	W	2	W	0	21	**15.6**
2813	*Industrial Gases*	W	17,854	0	7	W	W	0	0	3	**13.8**
2819	*Industrial Inorganic Chemicals, nec*	325	37,077	W	W	W	75	W	362	17	**8.5**
2821	*Plastics Materials and Resins*	636	14,780	668	192	210	W	1,074	0	W	**6.4**
2822	*Synthetic Rubber*	119	1,794	64	19	W	4,084	W	0	W	**14.1**
2823	*Cellulosic Manmade Fibers*	31	W	0	21	W	1	1,202	0	*	**25.3**
2824	*Organic Fibers, Noncellulosic*	W	6,976	W	53	W	W	W	0	1	**4.0**
2865	*Cyclic Crudes and Intermediates*	251	4,423	1,153	96	102	20,942	W	0	W	**12.1**
2869	*Industrial Organic Chemicals, nec*	2,705	15,104	1,747	502	W	W	3,819	0	784	**7.0**
2873	*Nitrogenous Fertilizers*	568	2,911	0	26	539	166	0	0	2	**23.1**
2874	*Phosphatic Fertilizers*	65	1,886	250	W	W	1	W	0	W	**4.9**

RSE Column (Natural Gas)
= 0.7
RSE Row (Chemicals and Allied Products)
= 6.1
Approximate RSE (Chemicals and Allied Products, Natural Gas)
= 6.1 · 0.7
= 4.3 percent
Approximate Standard Error (Chemicals and Allied Products, Natural Gas)
= (0.043) · (2,163) = 93 billion cu. ft.

Source: Energy Information Administration, Office of Energy Markets and End Use, Energy End Use and Integrated Statistics Division, 1991 Manufacturing Energy Consumption Survey.

The row and column factors are then computed by back-transforming the estimated model parameters; that is, by taking the $\log^{-1}$ of the model effects. This transformation yields

$$\begin{aligned} R_i &= \log^{-1}(\hat{m} + \hat{r}_i) = \log^{-1}\left(\overline{\log(RSE_{i,\bullet})}\right) \\ C_j &= \log^{-1}(\hat{c}_j) = \log^{-1}\left(\overline{\log(RSE_{\bullet,j})} - \overline{\log(RSE)}\right) \end{aligned} \tag{8}$$

For ease of presentation, the row factor includes the grand mean, m. Because of this factoring, the row factor for the ith row can alternately be expressed as the geometric mean of ith row:

$$R_i = \left(\prod_{j=1}^{J} RSE_{i,j}\right)^{\frac{1}{J}} \tag{9}$$

And, column factors, C_j, for a given table have a geometric mean equal to 1.0.

Since the MECS report presents a variety of energy-related estimates that are unique to certain industries, measures of the precision of population estimates are sometimes equal to zero or withheld from publication. When an RSE table contains a zero or withheld RSE, the table of RSE's is considered for generalization purposes to be "unbalanced". When the condition of an "unbalanced" table arises, substitute RSE estimates are inserted for these missing elements of the RSE table. Substitution of missing RSE's elements is based on an iterative procedure developed by Cochran and Cox.[43] A detailed description of the automated procedure used to produce the row and column factors appearing in this report can be found in Gargiullo and Goldberg.[44]

Sampling Error of Proportions

The estimates in this report can be used to produce proportion statistics based on the ratio of various estimates reported in the tables. Proportions are not given in the "Detailed Statistics Tables" but can be used to clarify the analysis. A proportion is the statistic of the form

$$\hat{P} = \frac{\hat{Y}}{\hat{X}}, \tag{10}$$

where $\hat{Y}$ and $\hat{X}$ are survey-based estimates of aggregate parameters Y and X, respectively, and characteristic X "encompasses" characteristic Y ($Y \subset X$). That is, each population element (and, thus, each sample case) that contributes to Y also contributes to X, and the value of X for each element is greater than or equal to the value of Y.

From standard errors given by Equation (**3**) that are then generalized by Equation (**4**), the approximate RSE's of aggregate statistics can be used to produce an upper-bound on the approximate errors for proportions. The straightforward additive error formula shown in Equation (**3**) gives rise to a similarly straightforward upper-bound approximation to the error of an estimated proportion. The approximation can be expressed in terms of the generalized RSE's of the aggregate statistics entering into the proportion as

$$\hat{RSE}(\hat{P}) \le \sqrt{\left[\hat{RSE}^2(\hat{Y}) \bullet (1 - 2 \bullet \hat{P})\right] + \hat{RSE}^2(\hat{X})}\,. \tag{11}$$

Justification for this formula is found in the MECS methodological report.

[43]Ibid.

[44]Gargiullo, P.M., and Goldberg, M.L. "A Modified Table Producing Language (TPL) for Producing Tables of Survey Statistics with Variances" *Proceedings of the Bureau of the Census Fifth Annual Research Conference*. (1989.)

Nonsampling Errors and Bias

Nonsampling errors that affect MECS sample units can be divided into four major categories:

- Operational errors, including editing, coding, and tabulation errors.
- Errors of measurement, including a lack of precision by the respondent, failure of the respondent to understand instructions, etc..
- Errors of estimation, including the assumptions underlying the derived values.
- Errors of nonobservation, including nonresponse and noncoverage.

These errors are collectively referred to as nonsampling errors because they are not related to the sampling process, and, thus, would be equally likely to occur in a complete census or a sample survey.

It is felt that operational errors are not a major concern for the estimates included in this report. The quality control procedures that were employed for check-in, editing, coding and keying the returned questionnaires (Appendix B) are standard procedures that are in place at the Bureau of the Census and have withstood the test of time. Data tabulations were independently verified by comparing marginal totals in tables generated from files supplied to EIA with corresponding totals generated directly from microdata files maintained at the Census Bureau.

Errors of measurement are a concern in any data collection activity. The survey results for the MECS were subjected to extensive computer editing procedures which were specifically designed to detect errors of measurement. Responses that failed these tests for reasonableness and consistency were recalled by analysts familiar with manufacturing processes and energy use. Major errors, including omissions and misreporting by orders of magnitude, were corrected. No editing procedure is capable of identifying all measurement errors, however, and some small errors will remain. To the extent that these errors are due to random, rather than systematic misjudgments, they are compensating in the aggregate totals presented in this report, and it is believed that there are few large systematic biases that result from them.

Errors of estimation of energy consumption could have resulted from the assumptions that underlie the derived values (see Appendix B), and the estimates of the consumption of onsite- and offsite-produced fuels and raw material inputs could be biased as a result of such errors. For example, the derivation logic makes the assumption that energy produced onsite at a manufacturing establishment is considered first as a shipped product, second as a feedstock, and lastly as a fuel. If that logic does not hold, derived estimate values will be misapportioned. However, considering the mechanisms required to produce energy onsite, it is highly probable that this logic accurately represents manufactures. These nonsampling errors, if present, are relevant only for tables in this report that are based on derived values. Estimates based upon reported values would not be subject to this potential source of bias.

Errors of estimation of energy consumption based on fuel-switching data could have resulted from the assumptions that underlie the formation of the maximum consumption estimates. Implicit in these estimates is the assumption that all potential switches at an establishment can be made simultaneously. For example, if a respondent indicates that natural gas could substitute for both distillate fuel oil and coal, the two quantities are summed together (after converting to like units) to contribute to the maximum consumption of natural gas. To the extent that one or more substitutions are constrained by the performance of another, the published maximum consumption quantities presented in Table A52 would overestimate the "true" value.

Finally, several potential sources of nonsampling error and bias result from errors of nonobservation. As described in Appendix B, the 1991 MECS represents, in terms of sampling coverage, the mail frame of the 1991 ASM or 98 percent of the manufacturing universe. Even though the MECS is a legislatively mandated survey and sampled establishments are given sufficient opportunity and time to respond, nonresponse occurs in the MECS and is accounted for in a nonresponse adjustment of sampling weights presented in Appendix B. Clearly, had these adjustments not been performed, the estimates produced from only the responding establishments would not have been representative of the target universe for the MECS. Such estimates would have been biased. Adjusting the sampling weights to reflect the target universe is an attempt to mitigate the potential effects of such a bias.

Adjustment factors are calculated for each of the 62 published strata to account for the variation of nonresponse between strata. Each stratum represents a relatively homogeneous subgrouping of establishments with respect to primary product output and level of fuel consumption. It is theorized that the MECS sampling procedure - selecting establishments based upon their relative amount of purchased electricity or fuel expenditures - would be reflected in adjustment factors using total energy costs (sum of fuel and electricity expenditures) as the control variable rather than using either purchased electricity or fuel expenditures.

Implicit in that procedure is the assumption that primary product output and level of fuel consumption are highly correlated with energy expenditure patterns, so that the establishments within a stratum would also be homogeneous with respect to the quantities, types, and shares of energy consumed as fuels and for nonfuel purposes. Also, the weight adjustment method assumes that the relationship between survey variables of interest and the control variable used for constructing the adjusted sample weight is the same for the population covered by MECS respondents within an adjustment stratum as it is for the rest of the population within that stratum.

To the extent that the nonresponding establishments within the adjustment stratum share the energy expenditure patterns of the responding establishments within the strata, the resulting adjustments to the MECS estimates will tend to be minimally biased. If, on the other hand, the energy expenditure patterns of the responding and nonresponding establishments differ substantially, the resulting adjustments are potentially biased, and the overall estimates may not accurately represent the originally targeted MECS universe.

More detailed information on sources of nonsampling error in the MECS can be found in the methodological report.