ASSESSING THE EFFECTS OF IMPUTED DATA ON SELECTED RESULTS FROM THE 1987 ECONOMIC CENSUSES

Leroy Bailey, Ann Jansto, and Charlene Smith, Bureau of Census
Charlene Smith, Room 3128, Bldg 4, Washington, DC 2023

KEY WORDS: Nonresponse; Telephone follow-up; Ratio estimation; Multiple imputation

The Bureau of Census conducts censuses of business establishments every five years. As in every census or survey, the 1987 Economic Censuses were subject to nonresponse, for which prescribed imputation procedures were followed. At the establishment level the procedures vary according to the census and item under consideration. The principal procedure involves the direct use of administrative data for missing census items. Also used is imputation based on previous census and survey data and estimates of period-to-period changes in the activity of the establishments or inter-item relationships. Administrative data refer to data for the designated census items that have been compiled from Internal Revenue Service (IRS) and Social Security Administration (SSA) files and stored in the Standard Statistical Establishment List (SSEL). The imputed data cited in the paper are derived from the SSEL, and reflect the results of imputation for item and establishment nonresponse and the editing or correction of data that were considered erroneously reported. These data are the values used in census publications.

An evaluation study was conducted to assess the accuracy of the imputed data for three of the 1987 Economic Censuses. This study will be referred to throughout this paper as the Evaluation of Imputed Data from the 1987 Economic Censuses, or EID. The EID focused only on selected standard industrial classification (SIC) groups within the Censuses of Wholesale Trade, Retail Trade and Service Industries. The four primary items of interest were first quarter employment (the number of paid employees on March 12, 1987), first quarter payroll, annual payroll, and sales and receipts (revenue for tax-exempt establishments). The secondary objectives of the study included efforts to facilitate the identification of misclassification problems encountered in the selected trade areas and to further the development of census imputation methodology. This paper discusses the survey design, the data collection methodology and the estimation procedure used for the evaluation and presents some of its results.

Sample Selection

A stratified systematic sample of about 3,000 establishments was selected for the EID. The principal stratification was based on type of unit (single or multi), trade area and SIC code. The 1987 SIC codes on which the evaluation is based are at the three-digit level for wholesale establishments and the two-digit level for retail trades and services. In addition, the SIC groups were stratified by categories based on establishment payroll to ensure that the precision of desired estimates was within an acceptable range.

The sampling frame for the study was the set of nonrespondents to the censuses as of August 31, 1988. Establishments were removed from the frame if they were not mailed a questionnaire, did not belong to any of the three trade areas selected for the study, or were inactive businesses. To ensure that the desired sample sizes were achieved for the smallest unit of analysis, the SIC group, adjustments were made to the respective sampling rates to allow for these contingencies and for the possibility of late respondents to the censuses.

Data Collection and Processing

Establishments in the sample for the EID were contacted initially by mail, reminded of the nonreceipt of their census form, and informed that they would be contacted by telephone and asked to respond to questions relating to several census items. The telephone interviews began on an average of five to ten working days after the introductory letters were mailed. All interviews were conducted from the Census Bureau's headquarters during the period from November, 1988 through July, 1989. After the requested data were collected from the sampled establishments, transcription reviews and preliminary edit checks were performed. Data which passed these checks were then compared to the corresponding administrative data. Establishments for which the ratio of the EID reported data to the administrative data fell outside of the range 0.5-2.0 were identified and investigated. Large discrepancies that could not be ascribed to data processing procedures were noted and reconciliation interviews were conducted. During these interviews, the interviewer probed for reasons why the reported data were different from administrative data, without informing respondents that the Census Bureau had current administrative data. About 10 percent of the sample required a reconciliation interview.

Interviewing performance statistics were collected weekly to monitor the progress of the data collection activities. Table 1 presents the distribution of interview outcomes.

Table 1. Distribution of Interview Outcomes for All Trade Areas Among Single Unit Establishments

Outcome Category	Number	Percent
Completed/Partial	2117	69.2
Refusal	330	10.8
Data Not Available	81	2.6
No Contact Made	494	16.1
Out-of-Scope	38	1.3
Total	3060	100.0

For all trade areas, about 69 percent of the single-unit interview attempts resulted in completed or partial interviews. About 11 percent of the establishments were refusals. The desired data were not available for roughly three percent of the cases, due to problems relating to such administrative matters as changes in ownership of businesses, establishment mergers or dissolutions and inept record keeping. Telephone numbers and addresses were not obtainable for approximately 16 percent of the establishments. A sizable portion of these cases were thought to have gone out of business, while others might have relocated or effected changes in their organizational structures and/or operations. This meant that for such cases the identifying information on the SSEL was not current, which encumbered efforts to contact them. Regarding other operational matters relating to the evaluation, it was observed that a large number (about 88 percent) of the successful interviews were conducted within ten minutes; however, it took two weeks or more to make productive contact with about 58 percent of these establishments. Tables 2 and 3 present the distribution of respondent establishments by call length and interview processing period, respectively.

Table 2. Distribution of Length of Calls Resulting in Completed/Partial Interview for All Trade Areas Among Single Unit Establishments

Length of Call	Number	Percent
1 - 5 minutes	1170	55.3
6 - 10 minutes	694	32.8
11 - 15 minutes	183	8.6
Over 15 minutes	70	3.3
Total	2117	100.0

Table 3. Distribution of Elapsed Time Between Initial Contact and Final Call for All Trade Areas Among Single Unit Establishments

Processing Period	Number	Percent
Same Day	316	14.9
1 week	574	27.1
2 - 4 weeks	683	32.3
Over 1 month	544	25.7
Total	2117	100.0

Following the data collection, the survey data were coded, keyed and verified, and the computer editing of the operational and reported data files occurred. These files were merged with a file containing the corresponding imputed data. This merge comprised the survey analysis file.

Estimation

For any census item of interest, y_{hij} will denote its value for the jth census nonrespondent of the ith payroll category and the hth SIC group. Let Π_{hij} be the selection probability for the establishment with this value. The value of y_{hij} reported in the evaluation study and the corresponding census impute will be given by $y_{hij}^{(r)}$ and $y_{hij}^{(c)}$ respectively. The number of payroll categories associated with the hth SIC group will be given by M_h; the number of nonrespondent establishments in the ith payroll group will be denoted N_{hi}, with the corresponding EID sample size given by n_{hi}. The estimator for assessing the accuracy of census imputation is the imputation correction ratio, designed to effect comparisons between item totals based on EID reported data and the corresponding totals from census imputes. This estimator at the SIC group level is denoted by the following equation:

$$\hat{R}_h = \frac{\sum_{i=1}^{M_h} \sum_{j=1}^{n_{hi}} y_{hij}^{(r)} \Pi_{hij}^{-1}}{\sum_{i=1}^{M_h} \sum_{j=1}^{n_{hi}} y_{hij}^{(c)} \Pi_{hij}^{-1}} = \frac{\hat{Y}_h^{(r)}}{\hat{Y}_h^{(c)}}$$

The estimator at the trade area level is derived by summing the $\hat{Y}_h^{(r)}$ and $\hat{Y}_h^{(c)}$ over all SIC groups and taking the ratio of the reported sum to the imputed sum. Clearly, if census imputes are exactly equal to the EID reported data, the ratio at either level will be 1.

Since the imputation correction ratio is a ratio of random variables, the variance can be approximated by the following:

$$Var(\hat{R}_h) \doteq (\hat{R}_h)^2 \left[\frac{Var(\hat{Y}_h^{(r)})}{(\hat{Y}_h^{(r)})^2} + \frac{Var(\hat{Y}_h^{(c)})}{(\hat{Y}_h^{(c)})^2} - \frac{2\,Cov(\hat{Y}_h^{(r)}, \hat{Y}_h^{(c)})}{\hat{Y}_h^{(r)} \hat{Y}_h^{(c)}} \right]$$

Since our sample consisted of independent systematic sampling within payroll categories, the variances of $\hat{Y}_h^{(r)}$ and $\hat{Y}_h^{(c)}$ and the covariance term can be approximated by weighted sums of sequential differences within each payroll category, summed over the payroll categories.

An establishment was defined to be a partial respondent if there was a response for at least one, but not all four, of the principal data items. If all four data items were missing, the establishment was denoted a total nonrespondent. Survey values for the missing items of partial respondents were

derived from linear regression models based on the responding establishments. Survey values for total nonrespondents were imputed through a hot deck based multiple imputation procedure. In this procedure, donors were drawn at random and with replacement from the same SIC group and payroll category as the nonrespondent. The donor's ratio of EID reported value to the corresponding census impute was then applied to the census impute for the nonrespondent to generate the nonrespondent's EID survey value. Imputation correction ratios at the SIC group and trade area levels were calculated from the resulting complete data set of EID respondents and the imputed values for the EID nonrespondents using the above equations. This process was repeated five times, generating five sets of imputation correction ratios which were then averaged. The results at the trade area level are presented in Table 4 along with standard errors and p-values based on the total variance from the multiple imputation procedure.

Table 4. Imputation Correction Ratio Estimates - Trade Area Level

	Wholesale	Retail	Services
First Quarter Employment	0.69 0.985 *0.035*	0.73 1.016 *0.044*	0.23 1.095 *0.069*
First Quarter Payroll	0.89 0.992 *0.055*	0.26 1.077 *0.067*	0.02 **1.233** *0.092*
Annual Payroll	0.59 1.015 *0.027*	0.12 1.159 *0.092*	0.04 **1.261** *0.122*
Sales and Receipts	0.42 1.075 *0.083*	0.18 1.141 *0.103*	0.06 **1.154** *0.074*

Analysis

Imputation Correction Ratios

Table 4 presents the imputation correction ratio estimates at the trade area level. Standard errors are given in italics in the lower right hand corner of each box, with p-values based on the hypothesis $\hat{R}$=1 in the upper right hand corner. Overall, the ratios indicate that census imputes are fairly accurate, with only three of the twelve trade area ratios exhibiting a p-value of 0.1 or less. However, it is interesting to note that ten of the twelve trade area level ratios and three-fourths of the SIC level ratios are greater than 1, suggesting a tendency toward underimputation in the censuses.

At the trade area level, imputation correction ratios for service items are among the most extreme (all three significant ratios are for service items), while the ratios for wholesale items exhibit the least amount of deviation from 1. At the SIC group level, ratios for all three trade areas vary tremendously between groups for each item. SIC group level ratios for the sales item within the wholesale trade span the greatest range of values, varying from 0.776 to 2.206. The second greatest range of ratio values occurs in the annual payroll item for services, with ratios ranging from 1.042 to 1.560. Although services and wholesale trade both exhibit large ranges of SIC group ratios for all four items, the ranges for wholesale ratios are centered around 1 while the centers of the ranges of values for service ratios are greater than 1. Of the 25 SIC group level ratios less than 1, 17 are within wholesale trade, 6 are within retail trade, and only 2 are within services. These differences are evident in the ratio values at the trade area level.

Establishment level ratios that were greater than 7.5 or less than 1/7.5 were considered outliers. These cutoffs were determined by observing the natural breaks in the data over all trade areas. One of the more interesting analyses of the EID has been the comparison of the distribution of establishments by imputation method for the outliers with the distribution by imputation method of all sampled establishments. The census impute for each item of an establishment is marked with a code indicating the method used, for the impute. Table 5 lists the imputation methods and codes. The vast majority of imputation method codes are "A", indicating the use of the administrative value. However, for the outliers, the percentage of establishments imputed with administrative values is much less than for all sampled establishments. Imputing with a zero ("Z") or a ratio adjustment based on 1987 industry averages ("J") is much more prevalent among the outliers, suggesting that these imputation methods might need further investigation. Table 6 presents the comparative distributions of source codes by item.

Table 5. Imputation Methods by Code

Code	Source Description
A	Administrative data
C	Corrected by problem solving clerk
D	Derived from other reported data
H	Ratio imputation based on 1982 census data for same establishment
I	Ratio imputation based on 1982 industry averages (cold deck)
J	Ratio imputation based on 1987 industry averages (warm deck)
M	Midpoint of range that will pass all edit checks
P	Ratio imputation based on prior year (1986) administrative data
R	Reported
Z	Zero imputed from blank

Table 6. Comparative Distributions of Establishments By Imputation Method and Item

Employment

Code	A	C	D	H	J	P	R	Z	m
All cases	2754	53	2	1	3	6	24	117	11
%	92.7	1.8	0.1	0.0	0.1	0.2	0.8	3.9	0.4
Outliers	13	1	0	0	0	2	0	22	2
%	32.5	2.5	0.0	0.0	0.0	5.0	0.0	55.0	5.0

First Quarter Payroll

Code	A	C	H	I	J	P	R	Z	m
All cases	2557	47	22	4	105	85	23	117	11
%	86.1	1.6	0.7	0.1	3.5	2.9	0.8	3.9	0.4
Outliers	12	1	1	0	7	5	4	15	2
%	25.5	2.1	2.1	0.0	14.9	10.6	8.5	31.9	4.2

Annual Payroll

Code	A	C	H	I	J	M	P	R	m
All cases	2537	55	23	3	161	5	149	27	11
%	85.4	1.9	0.7	0.1	5.4	0.2	5.0	0.9	0.4
Outliers	11	1	0	2	10	0	3	0	2
%	37.9	3.4	0.0	6.9	34.5	0.0	10.3	0.0	6.9

Sales and Receipts

Code	A	C	H	I	J	M	P	R	m
All cases	1426	200	201	21	624	2	459	27	10
%	48.0	6.7	6.7	0.7	21.0	0.0	15.5	0.9	0.3
Outliers	7	2	2	1	33	0	4	0	2
%	13.7	3.9	3.9	2.0	64.7	0.0	7.8	0.0	3.9

"m" indicates imputation code was missing

The remaining portion of this discussion will focus on service sales and receipts, to illustrate the additional analyses conducted on each ratio.

Figure 1 illustrates the imputation correction ratio estimates at the SIC group level. The bottom portion of each bar is the SIC level ratio, with the standard error on top. The thick horizontal bar indicates the value 1. Ratios at the SIC group level are both less than and greater than 1, indicating both over and under imputation. Eight of the nine ratios are greater than 1, which contributes to the trade area value of 1.154. Of these, the most significant is SIC group 86 - Membership Organizations - with a ratio of 1.22 and a p-value of 0.04.

Figure 1. Comparison of SIC Group Level Ratios For Service Sales and Receipts

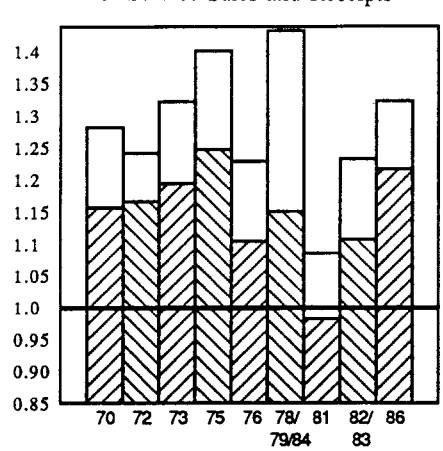
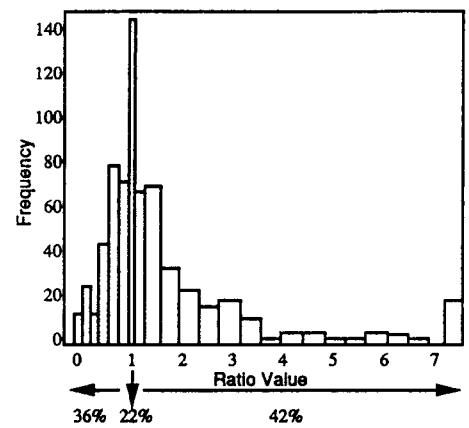


Figure 2 shows the distribution of establishment level ratios for respondents in service sales and receipts. For 22% of the establishments, the census impute was approximately equal to the EID reported value, yielding individual ratios between 0.99 and 1.01. For 42%, the census impute underestimated the reported value and for the remaining 36%, the census impute overestimated the reported value. Extremely high values and low values were capped at 7.5 and 1/7.5, respectively. These cutoffs were determined by observing the natural breaks in the data over all trade areas. For service sales and receipts, 24 establishments were "outliers", with 17 values greater than 7.5 and 7 values less than 1/7.5.

Figure 2. Distribution of Establishment Level Ratios For Service Sales and Receipts Respondents, Outliers Capped



Imputation correction ratio estimates were compared with similar ratios calculated with the outliers capped (at 7.5 and 1/7.5) and with respondents only. Capping the 24 outliers noted on the histogram had little effect on the services sales and receipts ratio, dropping it slightly from 1.15 to 1.14. The multiple imputation procedure also had little effect on the ratio, increasing it slightly from 1.13 based on respondents only to 1.15.

Misclassification Analysis

A secondary objective of the study was to assess misclassification rates for nonrespondents. Since item totals are published in the Economic Census reports by SIC group and trade area, they are potentially affected by misclassification. Verbal descriptions of the establishment's principal line of merchandise or service provided, obtained during the EID interview, were coded and verified manually. All ambiguous or unclear descriptions were coded to the SIC code present on the administrative file. Descriptions were coded to three digit SIC codes for wholesale and to two digit codes for retail and services. An establishment was considered to be misclassified if the imputed SIC code differed from the classification based on the reported data. Misclassification rates were then generated for the EID sample.

At the trade area level, misclassification was fairly consistent, at 11% for services and 13% for wholesale and retail. However, misclassification within trade areas, at the SIC group level, varied tremendously. For wholesale SIC

groups, misclassification rates, ranged from 3.9% for group 511/512, which includes paper products and drugs and drugstore supplies, to 47.8% for 504/508, which includes professional and commercial equipment, machinery and supplies. The misclassification rates for retail SIC group ranged from 6.9% for group 53/56, including general merchandise and apparel stores, to 24.7% for SIC group 57, home furnishings and equipment stores. For services, misclassification rates ranged from 0% for SIC group 81, legal services, to 35.2% for SIC group 73, business services.

Table 7 below presents the misclassification rates for all SIC groups.

Table 7. EID Misclassification Rates
By Trade Area and SIC Group

WHOLESALE		RETAIL		SERVICES	
SIC code	%	SIC code	%	SIC code	%
501	10.1	52	9.8	70	8.1
502	9.1	53/56	6.9	72	5.1
503/505	10.7	54	11.6	73	35.2
504/508	47.8	55	12.9	75	10.4
506/507	15.3	57	24.7	76	8.2
509	15.0	58	12.2	78/79/84	13.0
511/512	3.9	59	13.0	81	0.0
513	12.6			82/83	19.8
514/518	6.1			86	2.7
515/519	12.7				
516/517	7.4				

Conclusions

As stated earlier, the trade area imputation correction ratios are indications that at this level of aggregation, the imputation for the selected censuses is "reasonably" good for wholesale and retail trade. However, for three of the items included in the study, the trade area imputes for the service industries appear to underestimate the item value. While the wholesale and retail trade area estimates of the imputation correction ratios were not significantly different from 1.00, several seemed to have resulted from counteractions of considerably variable SIC level estimates. Tables 4 - 6 and the associated discussion of the previous section suggest a relationship between the use of imputation procedures, other than the substitution of administrative data, and the size of the corresponding imputation ratio adjustments. Specifically, increased use of procedures based on adjustments to alternative data sources to produce census imputes tends to generally produce increases in the size of the estimated imputation correction ratios.

Relative to census misclassification, a pattern was observed similar to that for the imputation correction ratios. At the trade area level the misclassification rates are between 11 and 13 percent; however, at the SIC level the corresponding rates vary considerably. In order to effect a reduction in the level of misclassification among census nonrespondents, perhaps census follow-up interview attempts, for which only data required to classify the establishment, could be considered.

Regarding potential imputation biases relating to differences in the effectiveness of the currently used imputation procedures, there is a need to seek alternative adjustment methodology appropriate for census items for which administrative data are more likely to be unavailable. Both theoretical and empirical research in this area seem warranted. Moreover, it is reasonable to consider the recurring estimation of correction ratios from follow-up efforts appended to census procedures or from analytical models applicable to relationships between respondents and nonrespondents in the populations of the censuses.