Imputation in Three Federal Statistical Agencies

John L. Eltinge, Bureau of Labor Statistics
Ralph A. Kozlow, Bureau of Economic Analysis
Donald M. Luery, U.S. Census Bureau

This paper has been prepared for presentation to the Federal Economic Statistics Advisory Committee (FESAC) on October 17, 2003. It represents work in progress and does not represent any agency's final positions on issues addressed. The FESAC is a Federal Advisory Committee sponsored jointly by the Bureau of Labor Statistics of the U.S. Department of Labor, and by the Bureau of Economic Analysis and the Bureau of the Census of the U.S. Department of Commerce.

Imputation in Three Federal Statistical Agencies

John L. Eltinge, Bureau of Labor Statistics
Ralph A. Kozlow, Bureau of Economic Analysis
Donald M. Luery, U.S. Census Bureau

Summary

This paper provides background on some of the imputation methods used by the Bureau of the Census (Census), the Bureau of Labor Statistics (BLS) and the Bureau of Economic Analysis (BEA) to adjust for incomplete data in establishment surveys. It discusses agency imputation procedures for establishments, enterprises, and other business organization units. In response to a request from FESAC, the material covered here is selected to provide a largely non-technical introduction. Section 1 provides a brief overview of concepts and methods, provides some literature references for more detailed developments of the underlying technical issues, and highlights some topics that one encounters frequently in imputation work for establishment surveys. Section 2 uses three examples from the BLS to display the wide variety of establishment survey settings in which nonresponse issues arise. These examples also illustrate some of the ways in which specific features of the surveys, and their primary nonresponse phenomena, lead to specific adjustment methods involving various combinations of weighting, deterministic imputation and stochastic imputation. Section 3 discusses some features of the StEPS system which provide a relatively flexible and unified framework through which most Census Bureau establishment surveys handle imputation work. Section 3 also highlights some special issues that arise when large units are nonrespondents, and when the underlying survey item may take either positive or negative values. Section 4 describes imputation work at the BEA for surveys of multinational companies (MNC's). Section 4 places special emphasis on adjustments for establishment births in one specialized survey, issues involving large nonrespondents, and imputation for gross and net flow data. Section 5 highlights some common practical considerations identified in Sections 2 through 4. Section 6 closes with a set of questions for FESAC.

1. Introduction

1.1 Overview of Concepts and Methods

In work with sample surveys, one generally encounters practical issues arising from nonresponse, i.e., the failure to obtain some or all of the requested data items for a given selected sample unit. The literature often distinguishes between *unit nonresponse* in which one obtains essentially none of the requested items from the sample unit, and *item nonresponse*, in which obtains some but not all of the requested items. Also, in panel surveys one often encounters *wave nonresponse*, in which a sample unit provides full or partial responses in one or more waves of the survey, but fails to provide responses in other waves. Nonresponse is of serious

practical concern because it may lead to biases and variance inflation in standard survey estimators. For some general background on survey nonresponse, see, e.g., Groves et al. (2002), Little and Rubin (2002) and references cited therein. To a substantial degree, the previous literature has tended to emphasize nonresponse in household or demographic surveys. Many of the general methodological ideas developed for household nonresponse apply in a straightforward way to establishment nonresponse. However, establishment nonresponse presents some special challenges arising from heterogeneity of the size of sample units, the availability in some cases of relatively rich auxiliary information (e.g., from the frame or other sources) for nonresponding units. Also, establishment nonresponse and household nonresponse arise from different underlying social, cultural and structural environments and response processes, which in turn may have implications for the types of auxiliary information that should be used in attempts to improve response rates and to adjust for the presence of nonresponse.

The current paper focuses on methods to adjust for nonresponse in establishment survey data. (Methods to reduce survey nonresponse may be covered in a subsequent FESAC paper.) In general, survey nonresponse adjustment methods include weighting adjustment and imputation.

Weighting adjustment modifies the customary probability weight associated with a given selected sample unit and is primarily used to adjust for unit nonresponse. In simple cases, weighting adjustment factors may be proportional to the estimated probability of response for the specified unit. In more complex cases, weighting adjustment factors are modified further to account for available auxiliary information, with the intention of reducing both bias and variance. Weighting adjustment is frequently used to adjust for unit nonresponse, but generally is not used for item nonresponse. To some degree, this arises from the fact that the costs of constructing and storing adjusted weights would become prohibitively large if one had a different set of adjusted weights for each potentially missing item. In addition, it can be problematic to incorporate information from partial responses into weighting adjustment procedures.

Imputation is a general term for the substitution of a specific numerical value for a given missing item. In *deterministic imputation* (sometimes called model-based imputation), the substituted value is obtained through a deterministic process; examples of the substituted values include the sample mean computed from a predetermined cell, or the predicted value obtained from a regression or ratio formula. In *stochastic imputation* the item is selected at random from the responses available in a specified cell, or group of sample units. Membership in imputation cells generally is based on auxiliary variables that available for both responding and nonresponding units. A given imputation cell is intended to be relatively homogeneous with respect to response probabilities, survey variables, or both. Some forms of imputation combine deterministic and stochastic elements; for example, some imputation methods replace a missing value with the sum of a model-based value from a regression or ratio expression plus a randomly selected residual term. In addition, some forms of weighting adjustment are mathematically equivalent to mean ratio or regression imputation. For example, if we use ratio imputation for item x_i using auxiliary variable y_i, and we use the weighted ratio $\hat{R} = \left(\sum_{resp} w_i x_i \Big/ \sum_{resp} w_i y_i \right)$ and the imputed value $\hat{x}_j = \hat{R} y_j$ then the estimate of the total using the imputes is $\hat{X} = \sum_{resp} w_i x_i + \sum_{nonresp} w_j \hat{x}_j$.

After some algebraic manipulation, this can be written as $\hat{X} = \sum_{resp} w_i^* x_i$ where $w_i^* = w_i \left(1 + \left(\sum_{nonresp} w_j y_j\right) \Big/ \left(\sum_{resp} w_j y_j\right)\right)$. If the auxiliary variable is the same for all items then there would just be a unique adjusted weight for a response unit. This can be done separately in each imputation cell. For this reason, "imputation" is sometimes used informally as a general term for any form of nonresponse adjustment.

In addition, some authors view all sample survey estimation as a type of incomplete-data problem. Under this framework, one may view standard survey estimation methods (e.g., probability weighting or calibration weighting) as forms of implicit or explicit imputation intended to account for nonsampled population units. See, e.g., Section 4.3 below and the literature on "designed missingness," in which sample units are randomly assigned to specific sections of a questionnaire. This broader view of imputation can be useful, although nonresponse work often is dominated by specialized issues that do not arise in probability sampling work. For example, nonresponse mechanisms generally are unknown and warrant careful data analysis, while probability sampling mechanisms generally are treated as known. Somewhat similar comments apply to adjustment for establishment births and deaths, which generally are handled through an adjustment step at an aggregate level, but which can in principle be viewed as part of the overall topic of incomplete data. Section 4.1 below provides additional discussion of birth-death adjustments for one specific case.

1.2 Preferred Properties of Nonresponse Adjustment Methods

Ideally, statistical agencies seek to use adjustment methods that have the following characteristics.

(1) The resulting point estimators are approximately unbiased.

(2) The resulting point estimators have variances that are minimized, to the extent possible in practice.

(3) One can use the resulting point estimator and associated measures of uncertainty to carry out valid inference for the underlying population parameters of principal interest.

In practice, one must evaluate criteria (1)-(3) within the context of specific models for the relationships among the response mechanisms, the outcome variables of interest, and available auxiliary information. For some general background on such models, and related controversies, see, e.g., Fay (1996), Rao (1996), Rubin (1996), Little and Rubin (2002) and the extensive references cited therein. A detailed discussion of model development and validation is beyond the scope of the current paper, but Section 6 includes some requests for comments regarding the use of model diagnostics and auxiliary variables.

Note also that characteristics (1)-(3) involve the combined effects of nonresponse, and related adjustments, on point estimation and inference for specific sets of population parameters. Thus,

for surveys that are focused on production of a small, well-defined set of parameters, it is useful to carry out diagnostic work directed toward characteristics (1)-(3) for those parameters; see, e.g., Rubin (1996, Section 1). However, for surveys in which a given imputation procedure may affect the estimators of a large number of parameters, direct assessment of characteristics (1)-(3) becomes more complicated, and agencies often focus attention on the properties of the individual imputed values.

2. Imputation for Establishment Surveys at the Bureau of Labor Statistics

Historically, work with nonresponse at the BLS arose from the general conceptual and methodological basis described in Section 1. However, specific imputation methods and related estimation work have tended to develop separately for individual BLS programs. To a large extent, this is attributable to specific features of individual surveys, e.g., their sample designs and periodicity, their predominant nonresponse patterns, the availability of relevant auxiliary data, the dimensionality of the data collected for a given unit, and the primary anticipated uses of the data. The following three examples illustrate some of the nonresponse adjustment concepts and methods commonly employed in BLS establishment surveys.

2.1 Job Openings and Labor Turnover Survey: Nearest Neighbor Imputation

The Job Openings and Labor Turnover Survey (JOLTS) is a monthly establishment survey carried out to obtain relatively rapid indications of changes in employment dynamics in the U.S. The sample design uses strata defined by the intersection of industry, geographical area and establishment-level employment size class. The selected sample units are asked to report six data elements: Total Employment, Total Number of Job Openings, Total Hires, Quits, Layoffs and Discharges, and Other Separations.

Crankshaw (2003, Chapters 8 and 9) provides some general background on nonresponse adjustments in JOLTS. For the current discussion, two nonresponse phenomena are of special interest. First, in conceptual and operational terms, the Total Employment variable is the simplest of the abovementioned data elements. Sample units that do not provide this element for a given month are unlikely to provide any of the other data elements, and thus are treated as unit nonrespondents for that month. To adjust for unit nonresponse, JOLTS uses weighting adjustment, with weighting cells defined by partially collapsed versions of the industry/area/size class strata described above.

Second, if a sample unit reports Total Employment for a given month, but does not report one or more of Total Number of Job Openings, Total Hires, Quits, Layoffs and Discharges, and Other Separations, then the result is viewed as item nonresponse and is addressed through a version of nearest neighbor imputation. For some general background on nearest neighbor imputation, see, e.g., Chen and Shao (1999), Fay (1999), Rancourt (1999) and references cited therein. For JOLTS, the nearest neighbor procedure is carried out within imputation cells. Within a specified

cell, a given item nonrespondent is matched with the responding unit that is closest to it, where distance is based on the Total Employment responses for the current month. The matched responding unit then serves as the "donor" of the ratios (missing element)/(Total Employment) that are missing from the item nonrespondent. Since the "donated" values are ratios rather than reported items themselves, the imputation procedure retains a somewhat greater degree of scale invariance than would be obtained through direct "donation" of the missing items themselves.

Properties of the JOLTS nonresponse adjustment procedures have been studied by several BLS staff members, but subject to availability of resources there are a number of additional areas of research that would be of interest. Examples include selection of optimal cell size and selection of optimal cell-collapse methodology, especially for cases involving very large units; adjustment of variance estimators to account more fully for cell collapse effects; and imputation alternatives involving multiple nearest neighbors.

2.2 Current Employment Survey: Implicit Imputation of Growth Rates and Pro-ration for Aggregated Reports

The Current Employment Statistics program (CES) is a large-scale monthly establishment survey carried out through a state-federal cooperative program. It collects and publishes information on employment, hours and earnings. Bureau of Labor Statistics (1997, Chapter 2; 2001, Chapter 1), Butani, Harter and Wolter (1997), Butani, Stamas and Brick (1997), Werking (1997) and West, Kratzke and Grden (1997) provide some general background on the Current Employment Survey.

The CES program recently underwent a transition to a probability sample design. This design uses extensive stratification by employment size class and industry within size class. For employment totals, point estimation is based on a "weighted link relative" estimator. Within a given estimation cell c, a relatively simple estimator of total employment for month t may be written as, $\hat{Y}_{tc} = X_{0c}(\prod_{m=1}^{t} \hat{R}_{mc})$ where X_{0C} is the total employment in this cell in a benchmark period 0 (known from administrative record data provided by the Covered Employment and Wages, or ES-202, program) and $\hat{R}_{mc}$ is a probability weighted ratio estimator of the growth rate in employment between months $m-1$ and m, respectively. Variance estimation is carried out through balanced repeated replication using Fay factors.

For the current discussion, two nonresponse issues are of principal interest. First, if a sample unit in cell c fails to respond in either month $m-1$ or month m, then it is excluded from calculation of the ratio $\hat{R}_{mc}$. Consequently, the weighted link relative estimator performs an implicit form of weighting adjustment within each cell c.

Second, the principal focus of the CES program is on production of estimates at the national and state level. However, many states have strong interest in production of estimates for sub-state regions, e.g., metropolitan statistical areas. For that purpose, it is important to note that although the CES data collection procedures request employment information at the worksite

level, some sample respondents are able to provide the information only at an aggregated level that may include several metropolitan areas. For such cases, local area estimation is based on "pro-ration" of the aggregated report to individual worksites, proportional to the most recently available ES-202 employment totals. Thus, this is a simple form of ratio-based imputation for the missing local information.

In conjunction with the probability redesign, the CES program has been the subject of a considerable amount of research in recent years. However, there are a number of possible areas for additional research in nonresponse adjustment for the CES. For example, as with JOLTS, one could consider development of additional diagnostics to guide the collapse of estimation cells that provide the basis for the weighting adjustment described above. Within this context, it may be useful to extend previously developed ideas of influence functions in sample surveys (e.g., Zaslavsky et al., 2001 and references cited therein) to evaluate the influence functions for specific observations within a given collapsed cell.

2.3 National Compensation Survey: Regression Imputation of Wage Rates and Nearest Neighbor Imputation of Benefits

The National Compensation Survey is a large-scale survey conducted by the BLS to collect information on wage and benefit data. The NCS uses a stratified multistage design with 154 primary sample units (PSUs) defined by geographical area. Additional stages of sampling include selection of establishments within selected PSUs, and selection of occupations within selected establishments. For some general background on the National Compensation Survey, see Bureau of Labor Statistics (1997, Chapter 8).

For the current discussion, four features of the NCS are of special interest. First, the NCS uses a five-year rotation sample design, where some of the selected sample units are asked to provide annual wage data, while other units are asked to provide wage and benefit data on a quarterly basis. Due to the rotation design, it is useful to distinguish between nonresponse that occurs at initiation (at which point relatively little information is available on the nonresponding unit) and nonresponse that occurs during update times (after successful completion of the initial data collection). Second, due to the multistage nature of the sample design, it is possible to have nonresponse at either the establishment or occupation levels of sampling. Third, due to inherent conceptual and operational complexities, benefit data tend to be more susceptible to nonresponse than wage data. Thus, with some rare exceptions one can view benefit nonresponse as being nested within wage nonresponse, in the sense that failure to provide a wage response tends to lead to failure to provide benefit responses as well. Fourth, there are complex multivariate relationships within a given vector of benefit data, and it can be important to preserve these relationships in the course of nonresponse adjustment.

In light of this, Barsky et al. (2000) focused attention on imputation for wage data that are missing during the initiation or update periods for data collection. Their results led to a recommendation for weighting adjustment to account for wage nonresponse at initiation. In

addition, for nonresponse in post-initiative periods, they recommended regression-based imputation using a model for the logarithm of the ratio of current wages to prior wages.

In addition, Buszuwski et al. (2003) recommended a combination of regression and nearest-neighbor imputation methods for missing benefit data, with somewhat different approaches at the initiation and update periods. Of special interest here are the different approaches used to impute three distinct types of benefits. First, imputed Social Security and Medicare premium payments can be computed directly reported wage payments, based the relevant legal requirements. Second, imputations for insurance and retirement benefits are more complex, but generally use nearest-neighbor methods. Third, imputed time-related benefits (e.g., annual leave or sick leave) are computed from an imputed value for the hours on leave, multiplied by the wage rate provided by the respondent.

Finally, as with JOLTS and CES, there has been a considerable amount of research on the NCS in recent years, but there are a number of additional research topics in nonresponse adjustment that are of potential interest. Barsky et al. (2000) and Buszuwski et al. (2003) provide some specific suggestions for such research.

Taken together, the examples in Section 2 suggest that no single approach to nonresponse adjustment can be expected to perform well across all establishment surveys. However, there are also substantial amounts of identifiable common structure for some groups of establishments, which in turn lead to some degree of commonality in methodological approaches to nonresponse adjustment. The Census Bureau has incorporated some of this common structure into the flexible imputation options provided in its Standard Economic Processing System (StEPS). The next section discusses these imputation options and describes some applications to specific establishment surveys at the Census Bureau.

3. Imputation in Economic Surveys at the U.S. Census Bureau

The Bureau of the Census designs and conducts monthly, quarterly, annual and quinquennial surveys to collect economic data that encompasses data for such areas as services, wholesale and retail trade, transportation, manufacturing, construction, financial statistics, research and development, medical insurance expenditures, capital expenditures, foreign trade and governments.

3.1 Imputation in the Standard Economic Processing System (StEPS)

Census has developed a general processing system called the Standard Economic Processing System (StEPS) for most of the surveys conducted by the Economic Directorate. This system has been developed to replace 16 legacy systems to provide a standard system and set of tools for data-collection support, editing, data review and correction, imputation, estimation, and system administration. The Annual Survey of Manufactures, surveys conducted by the Governments

Division, and foreign trade processing will continue to use other processing systems.[1] At this time, the following eight survey programs are using StEPS: Manufacturer's Shipments, Inventories, & Orders Survey (M3), Current Industrial Reports (CIR), Survey of Industrial Research and Development (R&D), Plant Capacity Utilization Survey (PCU), Manufacturing Energy Consumption Survey (MECS), Survey of Plant Capacity Utilization (PACE), Annual Survey of Manufactures – E-Commerce Business, Service Annual Surveys (SAS), Annual Retail Trade Survey (ARTS), and Annual Trade Survey (ATS)[2].

Imputation for StEPS is conducted in two modules: simple imputation and general imputation. Simple imputation imputes data considered equivalent to reported data and flags the resulting data as reported. A frequently performed type of simple imputation is data filling where StEPS fills in missing data that can be easily inferred from logical relationships among the data. A survey will use simple imputation when the change in the data will be small enough that it can still be considered the equivalent of reported data. For example, a set of detail items should add to a total item but one of the detail items is missing and the sum of the details is less than the total by a small amount. Simple imputation can set the missing detail to the difference between the total item and the sum of the detail items. Simple imputation flags imputed data as the equivalent of reported. Simple imputation is generally conducted before editing, review and correction. It is intended to ease the burden during these phases of the processing by automatically making minor corrections to the data.

The second module in StEPS for imputation is general imputation. This module is used after data review and correction to complete the data set by imputing for missing data and edit failures that could not be resolved during the data review and correction phase. General imputation also adjusts balance complexes so that detail items sum to total items. Interactive screens allow users to select from menus of methods for imputing individual items and from menus of actions for adjusting balance complexes. The general imputation module marks the changed values as imputed data.

StEPS has a module for estimation that can be used to reweight the sample to compensate for whole unit nonresponse. The discussion of imputation in this paper will be based on the StEPS General Imputation module. Simple imputation and reweighting the sample for nonresponse will not be discussed in this paper.

3.2 What is the methodology for estimating data for respondents who failed to respond or who partially respond?

StEPS uses the general imputation module for estimating the data for both whole unit nonresponse and for partial unit nonresponse by imputing for each individual item or adjusting a balance complex. This module divides the types of imputation into two categories: item imputation and adjustment of balance complexes. Each category has a number of different methods for imputation. Table 1 summarizes the methods for imputing individual item using the

1 The Annual Survey of Manufactures uses Economic Census Processing, and Foreign Trade Data Processing.
2 The 52 Current Industrial Report surveys are counted as one survey system as are the Annual Services Surveys.

following notation:

v = the item-name of the value being imputed
v' = the imputed value of v
z_j = the value of the j^{th} auxiliary variable
S(f) = the sum of item f over a defined set of records
$\left(S(f_1)/S(f_2)\right)_I$ = the ratio-of-identicals of items f_1 to f_2 where the numerator and denominator are both summed over the identical set of response cases.

Table 1. Methods for imputing individual items

Group	Name	Definition	Formula
Logical and Direct Substitution	SUM	Sum of auxiliary variables.	$v' = z_1 + z_2 + \ldots + z_n$
	RESIDUA	Auxiliary variable minus the sum of other auxiliary variables.	$v' = z_1 - (z_2 + \ldots + z_n)$
	PRODUCT	Product of two auxiliary variables.	$v' = z_1 \times z_2$
	VALUE	Value of the auxiliary variable.	$v' = z_1$
Mean	MEAN	Mean value of an auxiliary variable.	$v' = \overline{z}_1$
Ratio	RATIO	Ratio prediction for imputed item.	$v' = z_1\left(S(v)/S(z_1)\right)_I$
	ATREND	Auxiliary variable multiplied by a trend.	$v' = z_1(z_2 / z_3)$
	AUXRAT	Auxiliary variable times a ratio-of-identicals.	$v' = z_1\left(S(z_2)/S(z_3)\right)_I$
Regression	SIMPREG	Auxiliary variable times a regression coefficient.	$v' = b_1 z_1$
	MULTREG	Multiple regression prediction for imputed item.	$v' = b_1 z_1 + \ldots + b_n z_n$

In this table, an auxiliary variable may be the item to be imputed from another data collection period or another item from the same or a different data collection period. Except for the mean and ratio-of-identicals, the values for the auxiliary variables come from the unit to be imputed. General imputation calculates the mean and ratio-of-identicals based on all units that qualify in an associated imputation cell.

Table 2 shows the frequency in which each of the methods is used by survey. The most common methods are the ratio methods – ATREND, RATIO, and AUXRAT. ATREND uses a ratio based strictly on data from the unit being imputed. RATIO and AUXRAT use ratios calculated over all eligible units. AUXRAT is generalization of the RATIO method that provides more flexibility but requires a more detailed specification. RATIO is used in the simplest of situations when the numerator of the ratio is the item to be imputed and the denominator of the ratio is the auxiliary variable.

Table 2. Distribution of item imputation methods

		Item Imputation Methods								
SECTOR	**Primary Survey**	**ATREND**	**AUXRAT**	**PRODUCT**	**RATIO**	**RESIDUA**	**SIMPREG**	**SUM**	**VALUE**	**Grand Total**
Manufacture	ASMECB	0	0	2	0	0	0	0	2	4
	CIR	0	132	0	878	0	0	0	1269	2279
	M3	0	0	0	0	0	0	1	12	13
	PACE	0	72	0	0	0	0	0	0	72
	PCU	1	2	1	0	0	0	0	3	7
	RD	123	57	0	0	2	0	2	108	292
Manufacture Total		124	263	3	878	2	0	3	1394	2667
Service	ARTS	55	54	0	0	0	1	1	40	151
	ATS	12	26	0	0	1	1	0	19	59
	SAS	155	414	0	0	0	7	0	164	740
Service Total		222	494	0	0	1	9	1	223	950
Grand Total		346	757	3	878	3	9	4	1617	3617

StEPS allows the user to specify the methods to be used to impute an item. The specifications for imputing an item are characterized by the methods, the auxiliary variables, an ordering, and imputation conditions that determine when a specification may be used. The user may use the same method more than once using different auxiliary variables or different methods with the same or different auxiliary variables. The ordering determines the order in which each specification will be used and the imputation conditions determine if StEPS will try to use the specifications. If a specification has succeeded in imputing an item, general imputation skips the remaining specifications for that item. General imputation allows a great deal of latitude in deciding which methods will be used based on the available data. Tests on this availability would be included in the imputation condition for a specification. The available data can be from the same or previous data collection periods, or from external sources such as administrative records or the economic census.

The choice of methods depends on data that are available and the best ways to use these data to predict the missing value. Table 3 contains the imputation specification for the ARTS item *ectax00* (annual collected sales tax), involving the following items:

ecsal00 = unweighted annual sales, excluding sales tax
wcsal00 = weighted annual sales, excluding sales tax
etaxyn00 = indicator for sales tax collection: 1 for “yes”, 2 for “no”
wctaxy00 = recoded item that is equivalent to *wctax00* (weighted annual sales tax) when *etaxyn00* = 1 and is missing otherwise
wctaxb00 = recoded item that is equal to *wctax00* when *etaxn00* is in {1,2} and is missing otherwise

Table 3. General imputation specifications for ARTS item *ectax00*

Condition	Method	Formula	Auxiliary variables
etaxyn00=1	AUXRAT	$ecsal00 \times (S(wctaxy00) / S(wcsal00))_I$	$z_1=ecsal00$, $z_2=wctaxy00$, $z_3=wcsal00$
etaxyn00=missing	AUXRAT	$ecsal00 \times (S(wctaxb00) / S(wcsal00))_I$	$z_1=ecsal00$, $z_2=wctaxb00$, $z_3=wcsal00$

In records in which *ectax00* is marked for imputation and *etaxyn00*=1 (indicating collection of sales tax) the imputation of *ectax00* is based on a weighted ratio-of-identicals calculated from other records in the imputation cell that have *etaxyn00*=1. For records in which *ectax00* is marked for imputation and *etaxyn00* is missing, however, the imputation of *ectax00* is based on a weighted ratio-of-identicals calculated from records with either *etaxyn00*=1 or *etaxyn00*=2.

3.3 Large unit nonresponse

Census conducts extra and more intensive follow-ups for the large delinquent units than for the smaller ones. Surveys may have cutoffs in which the nonresponse units are followed up and those below the cutoff are not. The follow up procedures for the service area and the Construction Progress Reporting Surveys (CPRS) are discussed below.

For most of the data capture and review phases in services, the large and small delinquent cases are followed with equal intensity. Census' National Processing Center will initially fax the delinquent cases and then conduct follow-up interviews. When the deadline for ending the data capture nears, more emphasis will be placed on the larger cases. Much of this work will be done at the Census headquarters. Some surveys have 'must' cases from which a response must be obtained. Responses from the large multi-unit establishments are a high priority. If responses still cannot be obtained, Census attempts to find the requested data from other sources such as press releases, 10k annual filings to the SEC, stockholder reports, and any other source of information that can be found on the Web.

Some service surveys will run general imputation as frequently as daily. They will compare the imputed values for large cases for reasonableness and consistency with respect to similar companies in the same industry category and with the data (reported and imputed) for the case from the previous reporting period. For these cases, they will make further attempts to obtain data from the delinquent cases. (They also do this review for small and medium units in which the imputed value is a substantial percentage change from the previous reporting period.)

The CPRS during its monthly data collection cycle follow the small nonresponse units (construction projects under $100 million) and large nonresponse units (projects over $100 million) with the same intensity and the same procedures. They send follow-up forms, find better addresses and respondents for projects, and conduct telephone follow-ups. A construction project will take three or months to complete with the larger construction projects taking the most time. The CPRS will follow up monthly for up to three months for the smaller projects. If a smaller project has not responded by then, the CPRS classifies it as a permanent nonresponse and will make no

more attempts to contact it. For large projects, they will follow up, exhausting all means, possibly for up to twelve months at which time they will classify the nonresponder as a permanent nonresponse. The CPRS will impute for these cases.

Through the item imputation specification, users can create specifications so that larger nonresponse units will be imputed differently than small ones. First, the choice of a method specification to try may depend on the availability of certain data. To the extent that these data are more or less available for larger cases than smaller ones, these specifications will be used more or less frequently for the larger cases. Second, imputation conditions can be used to determine which specifications will be used for large units and which will be used for the other smaller units. For example in ARTS, some specifications are used for both Alphas (i.e., companies) and EINs (single and multiunit establishments), for Alphas only, and for EINs only. This is illustrated in Table 4 for current year (CY) sales. Third, separate imputation cells can be defined for large cases and smaller cases. For example, imputation for E-commerce sales in ARTS defines separate imputation cells for certainty cases (the largest cases at the time of selection) and for noncertainty cases.

Table 4. Imputation for Current Year Sales in ARTS

Condition	Method	Description
Both EIN and Alpha and reported CY Total Sales >0	AUXRAT	CY Total Sales × Ratio-of-identicals (CY Sales, CY Total Sales)
Alphas only and CY Annualized Sales > 0	VALUE	CY Annualized Sales
EINS only and CY Administrative Receipts > 0	VALUE	CY Administrative Receipts
EIN only and PY Sales > 0 and PY Payroll > 0	ATREND	PY Sales × (CY Payroll ' PY Payroll)
etc.		

3.4 Distribution of items using multiple methods

StEPS offers a special opportunity to examine the multiple uses of the imputation methods. The item imputation specifications are stored in General Imputation Specification (GIS) files, one for each survey. Table 5 shows by sector and survey, the number of items that use one, two, three, etc. method specifications. Predominantly, items have one or two specifications. However, the table is dominated by the Current Industrial Report (CIR) surveys that have 69 percent of the items and which have either one or two method specifications each. If we look at the other manufacturing surveys and the services surveys, predominately the items have two method specifications. For the Research and Development (RD) Survey, nearly all of the items have two specifications. For Services, most items have two specifications but about one-fifth have three specifications. Fifty-one of the items over all surveys have four or more specifications.

Table 5. Number of items using *n* number of methods

		Number of Methods									
SECTOR	Primary Survey	1	2	3	4	5	6	7	8	10	Grand Total
Manufacture	ASMECB	4	0	0	0	0	0	0	0	0	4
	CIR	259	1010	0	0	0	0	0	0	0	1269
	M3	1	6	0	0	0	0	0	0	0	7
	PACE	0	0	0	0	0	0	0	9	0	9
	PCU	0	2	1	0	0	0	0	0	0	3
	RD	11	102	4	9	0	3	0	0	0	140
Manufacture Total		286	1120	5	9	0	3	0	9	0	1432
Service	ARTS	5	44	8	2	2	0	0	2	0	63
	ATS	2	2	9	1	0	1	0	2	0	17
	SAS	19	220	58	10	0	6	3	0	1	317
Service Total		26	266	75	13	2	7	3	4	1	397
Grand Total		312	1386	80	22	2	10	3	13	1	1829

3.5 How does StEPS impute items that can be either positive or negative?

In the Quarterly Financial Report (QFR) survey, some of the items can have positive or negative values. The analysts review each and every filing for proper reporting, consistency with expectations based on other companies in the same present asset class, consistency with prior quarter reports, balance and omissions. As a first step, the analysts will contact the company in order to get a report from them. The next step would be to impute based on past filings. They will impute the value from the last quarter whether reported or imputed. This has the effect of carrying forward that last reported value of the company unchanged. Finally, the analyst would impute the average across the other companies in the same present asset class. For retail companies, analysts review these imputes for consistency for known seasonality of these data and make appropriate corrections.

Ensuring that balance complexes have both positive and negative details is a particularly difficult problem using the standard method for raking and can produce undesirable results. In the standard method of raking, raking forms a ratio by dividing the total by the sum of the reported details. Raking multiplies each detail by the ratio so that the revised details will now add to total. When details can be both positive and negative, the total can be much smaller or much larger in magnitude than sum of the details. This can either greatly shrink the details or expand them. Standard raking uses the formula $x_i' = x_i \left(y \Big/ \sum_j x_j \right) = x_i + x_i \left(R \Big/ \sum_j x_j \right)$ where y represents the total, the details before raking are represented x_i and x_j, R is the difference between the total and the sum of the details ($y - \Sigma_j x_j$), and the x_i' are the raked details. This practice has a sound statistical basis in situations in which the error in reporting a detail occurs at random and the variance of this error, var(x_i), is proportional to the value of the detail. The raked details are

optimal in the sense that they minimize the changes between the reported details subject to the raked details summing to the total. The measure of change that is minimized in the chi-square statistic

$$c^2 = \sum_i \frac{\left(x_i' - x_i\right)^2}{\mathrm{var}\left(x_i\right)}$$

When a detail can be negative, this assumption on the variance of the error is invalid. When the details are not restricted to being positive, Luery and Sigman (2000) show that if the variance is proportional to the absolute value of the detail then the adjustment formula

$$x_i' = x_i + x_i \left(R \Big/ \sum_j \left| x_j \right| \right)$$

finds the raked details that minimize the above chi-square statistic. The following example compares the results when using the standard raking method and the new method appropriate for positive and negative details.

Table 6. Example comparing the standard with the new method for raking

	Detail One	Detail Two	Sum of Details	Total
Initial Data	–100	200	100	120
Standard Method	–120	240	120	120
New Method	–93	213	120	120

4. Imputation at the Bureau of Economic Analysis:

The Bureau of Economic Analysis (BEA) designs and conducts about two dozen mandatory surveys of businesses that are owned by foreign companies or that own foreign companies, or that have purchased services from or sold services to unaffiliated foreign companies. These surveys may collect a substantial amount of accounting and operations data, such as data on balance of payments transactions or cross-border holdings, balance sheets and income statements, and data on imports and exports, R&D expenditures, employment and employee compensation, and sources and uses of funds. Some surveys are quarterly, annual, or quinquennial, and one survey (the survey of new foreign direct investment in the United States) is one time only.

4.1 Adjustments for unit births and deaths

In most cases, BEA makes no explicit adjustment to account for unreported births or deaths. To the extent that births or deaths involve a unit or a part of a larger affiliate, the matched sample ratios of period-to-period changes (which BEA applies to prior period reported or imputed data to impute current-period data) will reflect the impact of the birth or the death, and as a consequence, the birth or death will also have an impact on the current-period imputation of unreported transactions. In contrast, if the birth or death involves an entire affiliate, then that affiliate is excluded from the matched sample ratio described here, and therefore would have no impact on the imputation.

However, BEA does make an explicit adjustment for births in the case of form BE-13, Initial Report on a Foreign Person's Direct or Indirect Acquisition, Establishment, or Purchase of the Operating Assets, of a U.S. Business Enterprise, including Real Estate, and in the case of equity capital transactions reportable on its quarterly balance of payments surveys (forms BE-577 and BE-605). The following summary describes BEA's adjustment for births in the case of the "New Investment Form," form BE-13.

The BE-13 is a one-time survey, covering outlays by foreign direct investors or their existing U.S. affiliates to acquire or establish new U.S. affiliates. The featured statistic is total (gross) outlays to acquire or establish new affiliates, unreduced by selloffs or other transactions that reduce the level of foreign direct investment. Clearly, therefore, if even a single acquisition of a U.S. company were not reported, or was reported late, to BEA, the sum of reported data would be too low. The challenge is to devise a robust methodology for estimating unreported outlays, when unreported outlays are always a positive amount, but their magnitude may be highly volatile from year-to-year.

The preliminary (but not the final) BE-13 estimates include an imputation for survey reports that are received too late to be included in the published totals. The imputation is based primarily on the historical pattern of revisions and late reports affecting the series, with greater weight given to the pattern in recent years. That is, the increase in reported outlays, from the preliminary to the revised estimates covering recent years, is reviewed by subject matter specialists, who then judgementally set the "expansion" factor that is applied to the current year's preliminary data. The factor used here is intended to account for the effect of outliers in both the historical series and in the current year's reported data, outside information that may be relevant to the estimates for the current year, and changes in survey methods or procedures.

Separate adjustment factors are developed for real estate and for all other industries combined. Real estate is estimated separately because the impact of late reports tends to be larger in this industry than in the other industries. The adjustment factor for non-real-estate industries is usually between 5% and 15%, and the adjustment factor for real estate is between 20% and 60% (reflecting the large number of relatively small new investments in that industry which are reported late to BEA). In addition to total outlays, estimates are made for most other items obtained in the survey, such as of the total assets, sales, and employment of the newly acquired or established businesses. However, net income of nonrespondents is estimated as zero, because it is difficult to predict and, for new affiliates, often tends toward zero.

The aggregate estimates are carried down to individual country-industry cells. The cell-level adjustments are made to only the larger cells because it is difficult to estimate reliably data for small cells that typically have only a handful of investments in a given year. The threshold value for cells that are adjusted tends to change very infrequently.

4.2 Adjustments for nonresponding sample units, especially large units

As noted previously, BEA uses a matched sample ratio for estimating data for nonrespondents. The computer calculates this ratio at a detailed country-by-industry cell level. Large countries (Canada and the United Kingdom) have their own "expansion ratios" for a set of industries. Smaller contiguous countries (such as countries in Latin America, and in "other" Europe) are combined and a single expansion ratio for them as a group is calculated, and the ratio is then applied to the individual data item on the individual report form for businesses associated with that country or region.

The expansion ratios suggested by the computer program are often set within preestablished bounds (often + 30%) even if the matched sample data are outside of those bounds. This is because, if matched sample transactions are outside of these wide bounds, it is usually because one or two especially large transactions are having a major influence on the ratio. It is unlikely that unreported transactions would exhibit the same sharp percentage swings as the matched sample ratio, unless there is a widely prevalent influence on the data of which BEA should be aware. The computer-generated ratio is overriden by BEA estimators in cases where that seems warranted.

Data for both large and small respondents are handled by the computer in identical fashion. However, BEA does go to significant lengths to assure that the imputed data for large nonrespondents are highly accurate. In particular, BEA editors call every large nonrespondent, to obtain business cooperation in estimating the unreported data. Furthermore, BEA relies on publicly available information, including company SEC filings and news media reports, to develop its imputations. Also, editors carefully evaluate the reported or imputed data for the prior reporting period – which is used as the basis for the current period imputation – before deciding whether to accept the computer-generated imputation for the current period. In cases where the prior period imputed amount was unusually high or low due to a one-time or temporary factor, such as a labor strike, the editor will change the imputation for the current period to reflect more normal operations. These special efforts may only be undertaken for especially large affiliates and data, because a large amount of resources is required to investigate and evaluate individual imputations.

4.3 Comparison of adjustments for nonsample and nonresponding sample units

The same general methodology is applied by BEA to all imputed data, whether the imputations pertain to companies that should have reported but failed to do so, or to companies that were not required to report and thus did not do so.

In regard to the case of items that can be positive or negative, gross flows or positions are usually calculated first, and net flows are calculated as the difference between these gross amounts. More particularly, in the case of balance of payments transactions and positions, gross flows are calculated directly, except in the case of affiliate earnings. Thus, equity capital inflows and outflows, and receipts and payments of royalties and license fees, charges for other services, and interest, are separately calculated, and net flows are determined as the difference between the

gross amounts. In the case of earnings, however, the net amount is calculated directly. This is largely due to practical reasons – BEA does not collect information on gross operating revenues or expenses on a quarterly basis, and so cannot construct a matched sample for the gross flows.

There is a "wrinkle" that BEA must employ when imputing for amounts that were negative in the prior period. Take the example where the matched sample ratio shows a 20% increase; that is, the affiliates that earned $100 in the prior period earned $120 in the current period. If the ratio of 1.2 were applied to affiliates who sustained losses in the prior period, the resulting calculation of their earnings in the current period would show an incorrect direction of change. That is, affiliates who collectively lost $100 in the prior period would be shown in the current period as losing $120, if the matched sample ratio were used without adjustment. To avoid this counterintuitive result, the reciprocal of the ratio is calculated before it is applied to negative numbers in the prior period. Thus, in the above example, a ratio of 1 divided by 1.2 (or .83) would be applied to the firms that lost $100 in the prior period, and the imputation of their earnings in the current period would therefore be negative $83 (instead of negative $120).

4.4. If an agency has both a current period administrative record and a prior period report, which does it use in its imputation?

This is not applicable to BEA's imputation procedures. BEA does not have access to current period administrative records in estimating MNC data.

5. Closing Remarks: Some Common Practical Considerations for Imputation for Establishment Surveys

This paper has presented a largely nontechnical introduction to some of the general concepts, methods and complexities that arise in imputation and other nonresponse adjustment work with establishment surveys at the Census, BLS and BEA. Section 1 introduced the ideas of unit, wave and item nonresponse; discussed distinctions among weighting adjustment, deterministic imputation and stochastic imputation; and commented on some preferred properties of nonresponse adjustment methods. Sections 2, 3 and 4 highlighted some ways in which features of an establishment survey can affect decisions regarding the specific methods to be used to adjust for specific forms of nonresponse. Of special practical interest were the following.

Use of Auxiliary Data. In various forms, all of the imputation work considered here used auxiliary data in conjunction with explicit or implicit models. The choices of specific auxiliary data sources depended on several factors, including timely availability of the relevant data, and strength of association of the auxiliary data with the survey variables, the nonresponse mechanism, or both.

Use of Other External Information. In several cases, auxiliary microdata were supplemented with other forms of external information. Examples include the use of tax laws in calculation of

Social Security and Medicare payments in Section 2.3, and the use of special sources of public data like SEC filings in Sections 3.3 and 4.2.

Partial or Complete Preservation of Multivariate Structure. Much of the imputation work considered here was developed to ensure preservation of multivariate structure in the imputed record, to the extent possible. Examples include the imputation of ratios (rather than the missing items) in Section 2.1, linkage of leave-benefit costs with wage costs in Section 2.3, adjustment of balance complexes in Section 3.5, and adjustment of gross and net flow data in Section 4.3.

Nested Structure of Nonresponse. In many cases, definitional constraints or other factors produced a "nested" structure of nonresponse, in which failure to respond to a given item implied failure to respond to another item, at least with high probability. Examples included the employment count and separation data in Section 2.1, the wage and benefit data in Section 2.3, the sales and tax data in Section 3.2, the permanent nonresponse classification in the monthly CPRS in Section 3.3, and the flows data in Section 4.3. This nested structure in turn affects the amount of information available for use in imputation, and also affects the choice of imputation procedure to be used in a given case.

Large Units. Finally, there is especially strong interest nonresponse by, and imputation for, large establishments. Examples include selection of cell size in Section 2.1, diagnostics for collapse of cells in Section 2.2, follow-up and adjustment procedures for the CPRS and ARTS in Section 3.3, and follow-up procedures in Section 4.2.

In summary, qualitative characterizations of imputation procedures at the Census Bureau, BLS and BEA can identify important common factors, like the five listed above, but specific implementations of imputation procedures are also strongly affected by the characteristics of individual surveys, and individual groups of items within surveys.

Acknowledgements:

The authors thank James Burton, Shail Butani, Steve Cohen, Mark Crankshaw, Michael Davis, Larry Ernst, Vicky Garrett, Richard Hanks, Ned G. Howenstine, Larry Huff, Ron Lee, Marilyn Manser, Chester Ponikowski, Mark Sands, Richard Sigman, David Stringfield and Jean Swan for helpful discussions.

References:

Barsky, C., J. Buszuwski, L. Ernst, M. Lettau, M. Loewenstein, B. Pierce, C. Ponikowski, J. Smith and S. West (2000). Alternative Imputation Models for Wage Related Data Collected from Establishment Surveys. *Proceedings of the Second International Conference on Establishment Surveys (ICES II)*, 619-628.

Bureau of Labor Statistics (1997). *BLS Handbook of Methods.* Washington, DC: U.S.

Government Printing Office.

Bureau of Labor Statistics (2001). *CES Redesign Operating Manual (Interim)*. Technical document, Office of Employment and Unemployment Statistics.

Buszuwski, J.A, E.J. Elmore, L.R. Ernst, M.K. Lettau, L.G. Mason, S.P. Paben and C.H. Ponikowski (2003). Imputation of Benefit Related Data for the National Compensation Survey. Paper presented at the 2003 Joint Statistical Meetings, San Francisco, California.

Butani, S., Harter, R. and Wolter, K. (1997). Estimation Procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 523-528.

Butani, S., Stamas, G. and Brick, M. (1997). Sample Redesign for the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 517-522.

Chen, J. and J. Shao (1999). Jackknife Variance Estimation for Nearest Neighbor Imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Crankshaw, M. (2003). Technical documentation for the Job Openings and Labor Turnover Survey. Statistical Methods Staff, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association* **91**, 490-498.

Fay, R.E. (1999). Theory and Application of Nearest Neighbor Imputation in Census 2000. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Feenstra, R.C. and E.W. Diewert (2000). Imputation and Price Indexes: Theory and Evidence from the International Price Program.

Groves, R.M. and M.P. Couper (1998). *Nonresponse in Household Interview Surveys.* New York: Wiley.

Groves, R.M., D.A. Dillman, J.L. Eltinge and R.J.A. Little (eds.) (2002). *Survey Nonresponse.* New York: Wiley.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley.

Luery, D. and Sigman, R. (2000). Raking When the Details are Positive and Negative. Unpublished documentation, Washington DC: U.S. Census Bureau, Economic Statistical

Methods and Programming Division.

Rancourt, E. (1999). Estimation with Nearest Neighbor Imputation at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Rao, J.N.K. (1996). On Variance Estimation With Imputed Survey Data. *Journal of the American Statistical Association* **91**, 499-506.

Rubin, D.B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association* **91**, 473-489.

Sigman, R. (2001). Editing and Imputation in a Standard Economic Processing System. *Proceedings of Statistics Canada Symposium 2001 – Achieving Data Quality in a Statistical Agency: A Methodological Perspective.*

Werking, G. (1997). Overview of the CES Redesign. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 512-516.

West, S., Kratzke, D. and Grden, P. (1997). Estimators for Average Hourly Earnings and Average Weekly Hours for the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 529-534.

Zaslavsky, A.M., Schenker, N. and Belin, T.R. (2001). Downweighting Influential Clusters in Surveys: Application to the 1990 Post Enumeration Survey. *Journal of the American Statistical Association* **96**, 858-869.

6. Questions for FESAC:

1. Methods for formation of imputation cells. Some of the programs discussed in this paper obtain imputed values by sampling observed values from specified imputation cells. Some of these programs use a relatively large number of imputation cells, and decisions regarding formation and collapse of these cells may be data driven, to some extent. Please comment on specific cell-formation or cell-collapse methods, and related diagnostic tools, that you have found to be especially useful or problematic in imputation work with establishment data. Examples might include regression (for continuous survey variables), logistic regression (for response propensities), other parametric multivariate methods, CART (Classification And Regression Trees), and other nonparametric or semiparametric methods.

2. Methods for direct calculation of imputed values. Some programs use imputed values that are calculated directly from available auxiliary information, often through ratio, regression and related methods. For such cases, please comment on specific diagnostic tools that you have found to be especially useful or problematic in validation of the underlying models and assessment of the properties of a given proposed imputation method.

3. Use of administrative records from specific time periods. Agencies may have access to both an administrative record for the current period and to a prior period report. Please comment on which source should be primary for the imputation for the current period. If you believe that an answer to this question is likely to be very data-dependent, please describe some evaluation criteria that should be used in determining an appropriate question.

4. For surveys that use donor imputation, nonresponse variables may be imputed sequentially, individually or by groups of related items. Values of variables imputed previously would now be available as auxiliary variables for use in subsequent imputation work. Please comment on the advisability of basing imputations on variables that have previously been imputed, and the conditions under which this would be appropriate.

5. The Census Bureau, BLS and BEA have noted special challenges that arise in imputation for large units. Are there specific diagnostics (e.g., goodness-of-fit tests for models used in deterministic imputation, or evaluation of the properties of cells used in imputation, or evaluation of the distribution of values generated through stochastic imputation) that you consider especially useful in development and evaluation of imputation methods for large units?

6. Some of the literature on household surveys (e.g., Groves and Couper, 1998, Section 11.4) suggests that survey organizations “design for nonresponse” by collecting auxiliary data that may be useful in post-survey adjustment (e.g., weighting adjustment or

imputation) for nonresponse. To what extent, if any, are similar suggestions applicable to establishment survey nonresponse? Do you have any specific suggestions on ways in which to explore this through empirical studies?

7. As noted in the paper, some of the imputation literature has placed a strong emphasis on quantitative assessment of the relevant sources of error and their effects on published estimates. Please comment on specific tools (e.g., variance estimates, graphical displays, "fraction of missing information" diagnostics or more detailed sensitivity analyses) that you have found to be especially useful in communicating to various stakeholders (e.g., program managers, academic researchers and other data users) the components of variability (e.g., sampling, nonresponse, imputation and reporting error) associated with a given set of published estimates.

8. What information on imputation procedures should federal statistical agencies routinely publish? Examples might include general descriptions of imputation procedures; numerical information on imputation rates (e.g., aggregate imputation rates for each item, or more complex sensitivity analyses that display the effect of specific imputation methods on specific reported results), or general discussion of the limitations of imputed data. For each piece of information that you recommend publishing, what do you consider to be the appropriate forum, e.g., the summary press release, a technical appendix to the full published report, or supplements that are available on request but are not prominently featured in standard publications?