Overview of the Establishment Sampling Frame, Computer Edits, and Imputation Methodology for the 2002 Commodity Flow Survey

David L. Kinyon and Carol S. King
U.S. Census Bureau, Service Sector Statistics Division, Room 2651-3, Washington, D.C. 20233

Keywords: Establishment Surveys, Sampling Frame, Edits, Imputation, Commodity Flow Survey

1. Introduction

The Commodity Flow Survey (CFS) is conducted to produce estimates on the movement of goods shipped in the United States by manufacturing, mining, wholesale, and select retail establishments, as well as particular types of establishments that primarily provide company support. Estimates of shipment value, tons, and ton-miles are published by commodity, mode of transportation, shipment origin, and shipment destination. The survey is conducted as part of the quinquennial Economic Census through a partnership between the United States Census Bureau and Bureau of Transportation Statistics.

This paper[1] discusses various aspects of the 2002 CFS. Section 2 defines the desired "target population" and the achieved "sampled population" (Cochran, 1977). Section 3 discusses the construction of the establishment sampling frame. This paper does not discuss the sample design; however, see Black et al. (2003) for details on this topic. Section 4 gives an overview of the computer edits used to identify data that required review or imputation. Section 5 presents the methodology used to impute either value or weight for a shipment when the data item was missing or failed particular computer edits. Section 6 describes ongoing research for the 2007 CFS.

2. Target Population and Sampled Population

An *establishment* is the smallest business unit, which is usually a single physical location, where business transactions take place or services are performed. For the 2002 CFS, the *target population* was the set of establishments that were in business during 2002, were located in the United States, had paid employees, and would be classified as shippers in particular industries based on the 1997 North American Industry Classification System (NAICS) (U.S. Office of Management and Budget, 1998). These industries were the Manufacturing Sector (NAICS 31-33), Mining (Except Oil and Gas) Subsector (NAICS 212), Wholesale Sector (NAICS 42), Retail Electronic Shopping and Mail-Order Houses (NAICS 4541), as well as Warehousing (NAICS 493100) and Managing Offices (NAICS 551114) that primarily provided support to manufacturing, mining (except oil and gas), wholesale, and retail establishments of multi-establishment companies. Establishments that primarily perform support activities for their companies are called *auxiliary establishments*.

For the 1993 CFS and the 1997 CFS, the industries that defined the target population for each survey were classified according to the 1987 Standard Industrial Classification (SIC) system (U.S. Office of Management and Budget, 1987). Though an attempt was made to maintain similar industry coverage for the 2002 CFS, there were some changes in industry coverage due to the conversion from the SIC system to the NAICS. Most notably, coverage of the logging industry changed from an in-scope Manufacturing SIC code (SIC 2411) to an out-of-scope Agriculture, Forestry, Fishing, and Hunting NAICS code (NAICS 1133). Also, coverage of the publishing industry changed from in-scope Manufacturing SIC codes (SIC 2711, 2721, 2731, 2741, and part of 2771) to out-of-scope Information NAICS codes (NAICS 5111 and 51223). Although there was some loss in comparability between the 2002 CFS estimates and those from prior surveys, we felt it was important to maintain comparability with estimates from future CFS surveys.

The target population for 2002 was comprised of two types of establishments. The first type consisted of single-establishment companies, which were referred to as *single-units*. The second type consisted of establishments from companies that were comprised of multiple establishments, and these establishments were referred to as *multi-units*. All auxiliary establishments were multi-units.

The *sampled population* for 2002 was different from the target population because of the manner in which the 2002 CFS was conducted. Between January 6, 2002 and January 4, 2003, respondents to the survey were asked to report data for a sample of their shipments in each quarter of the year. A respondent was asked to report for assigned weeks that were in the same relative position within the quarters. Because we hoped to obtain improved response by mailing the survey forms

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

prior to the reporting weeks, the sample had to be selected in December 2001. Because of the time needed to construct the sampling frame, our sampled population could consist of only the establishments that were in business as of early September 2001, excluding establishments that came into existence at the end of 2001 or during 2002. Instead of sampling these new establishments at a later time, we performed an industry-level weighting adjustment to the 2002 CFS estimates to account for this activity, using preliminary data from the 2002 Economic Census. For more information on the census-adjustment procedures used in estimation, see Evans and Cantwell (1995).

3. The Establishment Sampling Frame

The following subsections discuss the establishment sampling frame in more detail. Section 3.1 describes the inputs to the sampling frame. Section 3.2 discusses the criteria that were used to determine the records included on the sampling frame. Section 3.3 describes the methods that were used to determine a measure of size (MOS) and a geography code for each record on the sampling frame.

3.1 Inputs to the Establishment Sampling Frame

There were two types of data sets that were input to the establishment sampling frame. The first type contained establishment data on industry classification, size, and geographic location from different sources. These sources included the Census Bureau's Business Register, which was called the Standard Statistical Establishment List (SSEL) at the time, the 1997 Economic Census, the 1999 Annual Survey of Manufactures (ASM), and the 1997 CFS. The second type of input contained information that was not specific to particular establishments, but could be used to determine an establishment's industry classification, size, and geographic location.

The Business Register is a database of all known establishments located in the United States or its territories. This database is periodically updated with administrative data from other government agencies, including the Internal Revenue Service (IRS), the Social Security Administration (SSA), and the Bureau of Labor Statistics (BLS). These administrative data include name and address information, industry classification codes, quarterly payroll data, and annual receipts data. The Business Register is also periodically updated with annual data at the establishment level from the Economic Census, ASM, and Company Organization Survey (COS).

The administrative data are provided by Employer Identification Number (EIN), which the IRS uses to identify business units that are aggregations of one or more establishments. For a given single-unit, the company, EIN, and establishment all refer to the same business unit. However, companies that own multiple establishments may use one or more EINs, and each EIN may consist of one or more multi-units.

The data extracted from the SSEL were based on the final SSEL data sets for 2000, which were created in September 2001. Because imputation for missing 2000 payroll had just been performed for multi-units, the 2000 payroll data were available. As will be discussed in Sections 3.2 and 3.3, these payroll data were used to determine the records included on the sampling frame and to calculate the MOS for each establishment.

Data were extracted from the 1997 Economic Census and the 1999 ASM databases. These data included NAICS codes and data items that could be used to create a measure of an establishment's size. The extraction of the 1997 Census data was necessary because the SSEL stored only a subset of the data items from the Census. The 1999 ASM data were used because the data from the 2000 ASM had not yet been finalized for publication.

We created a data set that contained information on the 1997 CFS establishments. It included estimates of 1997 value of shipments for establishments that contributed to the estimates. The data set also included a code that indicated whether a given managing office had been classified as a nonshipper.

Three additional data sets were input to the sampling frame and were used to determine an establishment's industry, size, and geographic location. The first data set mapped an SIC code to its most likely NAICS code, based on the 1997 Census distribution of NAICS codes for the SIC code. The second data set contained estimated coefficients from regression models that attempted to approximate CFS value-of-shipments data from administrative payroll and receipts data. (Section 3.3 gives more information on the regression methodology.) The third data set was used to assign an establishment's Metropolitan Area (MA) from the SSEL's geography codes.

3.2 Criteria for Inclusion on the Establishment Sampling Frame

The criteria that were used to determine the records included on the establishment sampling frame attempted to cover the sampled population described in Section 2, using available information. Each establishment had positive 2000 payroll and was located in the United States, based on its state geography code. Each single-unit EIN was active on the IRS's Master File, which meant that employee payroll withholdings were being reported to the IRS using Form 941: *Employer's Quarterly Federal Tax Return*. Each multi-unit was active on the SSEL, based primarily on results of the

COS.

We included records on the sampling frame, based on NAICS classification and the SSEL's type-of-operation code (TOC). The TOC was used to identify particular types of wholesale and auxiliary establishments. Because the SSEL was still on an SIC basis, we assigned a NAICS code to each establishment record, based on the method that was planned for the 2002 Census. In a few cases, we resolved differences between the NAICS code and TOC, using information on type of operation from the 1997 Census.

For the assignment of NAICS codes, we used NAICS codes from the 1999 ASM or 1997 Census for about 78% of the records. For almost all of the remaining records on the sampling frame, we mapped the SIC code on the SSEL to the SIC's most likely NAICS code, based on the 1997 Census distribution of NAICS codes for the SIC code. However, for a few single-units having a blank SIC code, we used the best available NAICS code from administrative data.

Table 1: Frequencies of Records on the Establishment Sampling Frame by Industry and Establishment Type

Industry	**Single-Unit**	**Multi-Unit**
Wholesale	272,403	98,570
Manufacturing	265,906	66,197
Retail Electronic Shopping & Mail-Order Houses	8,879	1,522
Other Retail	7,641	6,660
Managing Offices	NA[2]	18,473
Mining	3,792	3,088
Warehouses	NA	4,390
Total	558,621	198,900

Table 1 gives frequencies of the 757,521 records on the establishment sampling frame by NAICS-based industry classification and establishment type. Almost 74% of the records were for single-units. Combined, wholesale and manufacturing made up about 96% of the single-unit records and about 83% of the multi-unit records.

Only 10,401 of the records classified as retail on the establishment sampling frame were for electronic shopping or mail-order houses. To improve coverage of the Wholesale Sector, we included on the sampling frame 14,301 records for establishments that were classified in particular retail industries (automotive parts and accessories, tires, floor coverings, building materials, nursery and garden, and office supplies) in the 1997 Census and had indicated 0% sales to the general public in the 1997 Census. These establishments were likely to be classified as wholesale in the 2002 Census. Of the establishments selected for the 2002 CFS from this set of establishments, only those that were classified as wholesale in the 2002 Census were included in the estimates for the final report.

We attempted to prevent overcoverage of auxiliary establishments. We removed records for managing offices that had been classified as nonshippers in the 1997 CFS, because we believed that a managing office that did not have shipments in 1997 was also likely to not have shipments in 2002. We also removed records from the sampling frame for auxiliary establishments in two ways, based on primary industry served, because an auxiliary establishment was assigned a NAICS code based on its primary function, instead of its primary industry served. First, we removed the records for auxiliary establishments if they could not be matched within their companies to other establishments that were either on the sampling frame or classified in the Retail Sector (NAICS 44-45). Second, we removed the records for auxiliary establishments if their SIC codes, which reflected primary industry served, were not considered industries within the scope of CFS.

We considered performing an additional operation, which would have further limited the number of records for managing offices on the 2002 CFS sampling frame. For the 1997 CFS sampling frame, the number of records for managing offices was reduced to a large extent, based on the results of the 1992 Census. The record for a managing office was included on the sampling frame only if the presence of sales or end-of-year inventories had been indicated in the 1992 Census. However, research conducted prior to the construction of the 2002 CFS sampling frame showed that not all managing offices with shipping activity in the 1997 CFS had indicated sales or end-of-year inventories in the 1997 Census. Therefore, the 1997 Census results were not used to limit the number of records for managing offices on the 2002 CFS sampling frame.

3.3 Establishment Size and Geography Code

For each record on the 2002 CFS sampling frame, we assigned a measure of the establishment's size and a geography code. The MOS was used in the sample design, edits, and adjustment of estimates. The geography code was used in the sample design and was based on pieces of the U.S. states and the District of Columbia.

[2] In this paper, "NA" stands for "Not Applicable."

In creating the MOS for a given establishment, we attempted to estimate the establishment's annual value of shipments, which we hoped would be highly correlated with its 2002 CFS value and weight data. The inputs to an establishment's MOS included:

1. Estimates of 1997 value of shipments, adjusted using preliminary results of the 1997 Census, for establishments that contributed to the 1997 CFS estimates.
2. Proxies for CFS value of shipments from the 1997 Census and the 1999 ASM. Based on analyses of the 1993 and 1997 CFS, there did not appear to be suitable Census proxies for auxiliary establishments.
3. Payroll data for 1997, 1999, and 2000 from the 1997 Census, the 1999 ASM, and the SSEL.
4. Single-unit administrative receipts for 1999 from the SSEL.
5. Estimated coefficients from regression models that attempted to approximate CFS value-of-shipments data from administrative payroll and receipts data. Note that, because of the time required to review these estimated coefficients, they were computed from a preliminary sampling frame, for which the 1999 payroll data were the most recent available. The procedure was performed separately for single-units and multi-units by NAICS code and limited the influence of outlying observations (Mosteller and Tukey, 1977).

We wanted the MOS to reflect a full year of activity for 2000, based on its incorporation of the 2000 payroll data in payroll-based inflation factors or in the regression models. If there were indications on the SSEL that an establishment's 2000 payroll did not represent a full year of activity, we inflated the data. Also, we adjusted the 2000 payroll data using edits that compared these data to employment data.

The formula used to calculate the MOS for a given establishment depended on the establishment's NAICS code, TOC, type of establishment (i.e., single-unit or multi-unit), and the data that were available for the establishment. For the establishment sampling frame, Table 2 gives frequencies of its records by primary MOS input and type of establishment. Proxies from the ASM or Census were generally preferred and used for about 77% of the single-units and 86% of the nonauxiliary multi-units. For about 97% of the auxiliary establishments, an estimated regression coefficient was applied to the 2000 payroll data.

Table 2: Frequencies of Records on the Establishment Sampling Frame by Primary Measure-of-Size Input and Type of Establishment

Primary Input to MOS	**Single-Unit**	**Multi-Unit**	
		Auxiliary	**Other**
1999 ASM Proxy	23,794	NA	32,709
1997 Census Proxy	405,937	NA	118,168
1997 CFS Value of Shipments	51	667	54
Payroll-Based Regression	69,557	22,196	25,106
Receipts-Based Regression	59,282	NA	NA
Total	558,621	22,863	176,037

The primary geographic units used for the 2002 CFS were *state pieces* that were determined by the location of the 273 MAs in the United States. The MAs were a combination of the Metropolitan Statistical Areas (MSAs) and the Consolidated Metropolitan Statistical Areas (CMSAs), based on population counts from U.S. Census 1990. MA definitions based on U.S. Census 2000 were not yet available when we constructed the 2002 CFS sampling frame. For more information on MA definitions, see U.S. General Accounting Office (2004).

A given state referred to one of the fifty U.S. states or the District of Columbia. State pieces were determined by the intersection of the states with the MAs. For example, the Louisville MA intersected two states, resulting in two state pieces - the Louisville, KY piece and the Louisville, IN piece. The piece of a given state that was not associated with an MA was called its *Rest of State (ROS)*. The collection of all state pieces formed a partition of the United States.

We created the MA code based on the geography codes for each establishment on the SSEL. We matched each establishment's state, county, and place codes to the input data set that mapped these codes to the corresponding MA code. The MA code was based on U.S. Census 1990 and indicated either a state's MA or its ROS.

4. Computer Edits

After the sampling frame had been constructed and the sample had been selected, we established computer edits on the data, which included data on 2,649,761 reported shipments. The computer edits were run at the Census Bureau Headquarters in Suitland, MD after the data from the 2002 CFS forms had been screened, keyed, and transmitted by the National Processing Center (NPC) in Jeffersonville, IN. The screening procedure consisted of a set of manual edits that attempted to ensure completeness and accuracy of the data, including the units of measure that were used to report shipment value and weight.

The computer edits consisted of weekly follow-up edits, end-of-quarter aggregate tabulations, quarter-to-quarter consistency edits, and imputation edits. The four types of computer edits are described in more detail in the following subsections. The first three types were used to identify data that required review, while the last type was used to determine instances in which imputation would be performed. Because the imputation edits depended on the weekly follow-up edits, both types of edits were run prior to imputation.

Two types of parameters were used in the edits, with the goal of identifying outliers that could be reviewed using available resources. Before the edits were run, edit parameters that depended on reported shipment data were updated, after restoring originally reported data that had been overwritten during prior edit or imputation runs. Initially, reported data from the 1997 CFS were used for these edit parameters. We also used fixed edit parameters, which were adjusted after analyzing the edit results.

4.1 Weekly Follow-Up Edits

Batches of keyed data from the 2002 CFS forms were transmitted on a weekly basis by NPC to the Census Bureau Headquarters. Beginning in April 2002, weekly follow-up edits were applied to each batch. These edits were used to identify data that required follow up by the NPC clerks and Census Bureau analysts, and they consisted of three types. There were three sampling-related edits, four edits on missing or invalid shipment data, and six consistency edits on shipment data.

For each quarter of the survey year, respondents to the 2002 CFS were asked to report data for a sample of their shipments that were made during their assigned reporting week. We designed the sampling-related edits to identify potential problems with a respondent's sampling. The first edit compared the reported total number of shipments for the week to both the number of reported shipments and the number of reported shipments that contained value data. The second edit compared the number of reported shipments to the expected number of reported shipments. The third edit weighted the value-of-shipments data on the form and compared this estimate to the establishment's MOS.

We created the edits on missing or invalid shipment data to target data items that were critical to the published estimates. Shipments that failed these edits could have been the result of reporting errors or survey processing errors. The first edit identified shipments that were not exports and had missing or invalid data on U.S. destination. The second edit identified missing or invalid shipment commodity codes, based on the Standard Classification of Transported Goods (SCTG) commodity codes. The third edit identified shipments with missing or invalid mode of transport. The fourth edit identified shipments for which both value and weight were missing or zero.

The first consistency edit on shipment data compared the first two digits of a shipment's SCTG and the first three or four digits of its establishment's NAICS code to a table of valid combinations that was prepared by Census Bureau analysts. Commodities and industries that were often heterogeneous, such as SCTG 43 (Mixed Freight) and NAICS 339 (Miscellaneous Manufacturing), were excluded from this edit. About 9% of the shipments failed this edit.

For commodities in which shipment weight is often reported in tons, instead of pounds, the second consistency edit was a ratio edit that attempted to correct weight data that were not in the correct units of measure. In this edit, a shipment's weight-to-value ratio was compared to the median weight-to-value ratio for the shipment's commodity, which was calculated using all the shipments with reported positive value and weight data in the SCTG code. This edit corrected weight data for only a small number of shipments, which made up less than 1% of the total number of shipments.

The third consistency edit, which was called the Value-to-Weight Edit, attempted to identify shipments having unusual value-to-weight ratios for their commodities. This ratio edit compared the natural log of a shipment's value-to-weight ratio to the upper and lower bounds for its SCTG code. The natural log transformation was used so that the resulting outliers would be less affected by shipments with large ratios and would be more symmetric with regard to the tails of the distribution. The upper and lower bounds for the SCTG code were determined by calculating the transformed ratio for each shipment with positive value and weight data in the SCTG code, calculating the quartiles of the distribution of transformed ratios, and adding or subtracting three times the interquartile range to the median. This edit also identified shipments for which value was zero and weight was positive, or vice versa. Less than 5% of the shipments failed this edit.

The last three consistency edits targeted potential

problems with a shipment's mode of transport or hazardous materials code. The fourth consistency edit identified commodities whose mode of transport should not have been pipeline. The fifth consistency edit identified weight data that fell outside either a fixed upper or lower bound for shipments transported by parcel, truck, or air. The sixth consistency edit compared a shipment's hazardous materials code to a table of valid codes that was prepared by Census Bureau analysts.

Edit failures for all three types of weekly follow-up edits were sorted by batch and size, printed, and sent to the NPC clerks or Census Bureau analysts for follow up. The responsibility for following up the sampling-related edit failures was split. Failures for the first edit were followed up by the NPC clerks, while the other failures were followed up by the Census Bureau analysts. All other weekly follow-up edits were followed up by the NPC clerks.

4.2 End-of-Quarter Aggregate Tabulations

The end-of-quarter aggregate tabulations identified establishments whose contributions to the 2002 CFS estimates should be reviewed. The tabulations consisted of establishment contributions to key estimates of shipment value and weight, as well as establishment comparisons of the weighted estimate of annual value of shipments to the MOS. We started running these edits in July 2002, and we ran the edits after imputation of shipment value and weight data had been run. (For information on the imputation of these data, see Section 5.) We typically ran the imputation after the Census Bureau analysts had determined that around 90-95% of the expected response for a given quarter had been achieved. For a given quarter, we expected to receive about 70% of the forms that were mailed.

The first set of end-of-quarter aggregate tabulations identified establishments based on their percent contributions to weighted estimates of annual shipment value and weight by key tabulation cells. The cells were defined by 2-digit SCTG code and state, as well as by single mode of transport and state. The largest and smallest five establishments by percent contribution in each cell were identified, and data for these establishments were printed for the Census Bureau analysts to review.

The second set of end-of-quarter aggregate tabulations identified establishments based on the difference between an establishment's weighted estimate of annual value of shipments and its MOS. Establishments having a difference greater than $0.5 billion in absolute value or having one of the measures greater than 10 times the other were identified, and data on these establishments were printed for the Census Bureau analysts to review.

4.3 Quarter-to-Quarter Consistency Edits

The quarter-to-quarter consistency edits identified establishments whose data between successive quarters should be reviewed. For a given establishment, we compared data for the first and second quarters, the second and third quarters, and the third and fourth quarters. In July 2002, we started running these edits at the same time as the end-of-quarter aggregate tabulations, after a sufficient number of second quarter forms had been processed and subjected to weekly follow-up.

The first quarter-to-quarter consistency edit identified establishments for which the reported total number of shipments for one of the quarters was more than 10 times the number for the other quarter. Because we wanted to target establishments for which the reported total number of shipments varied greatly by quarter, we restricted this edit to establishments with at least 50 reported total number of shipments in each quarter being compared. The top 200 establishment failures by size were printed for the Census Bureau analysts to review.

The second quarter-to-quarter consistency edit identified establishments for which the average shipment value-to-weight ratio was more than 10 times the average for the other quarter. This edit attempted to identify establishments for which different units of measure were used or different types of shipments were reported from one quarter to the next. The top 200 establishment failures by size were printed for the Census Bureau analysts to review.

4.4 Imputation Edits

The imputation edits identified shipment value or weight data that required imputation. For a given shipment's value and weight data, we assigned corresponding Report/Impute (R/I) codes during data entry or the edits to indicate the data source or to flag a data item for imputation. If a Census Bureau analyst or NPC clerk corrected data, verified data, or flagged one of the two data items for imputation, the imputation edits did not overwrite the R/I code.

If a given shipment failed the Value-to-Weight Edit described in Section 4.1, an imputation edit was performed to determine which of the two data items would be set for imputation. If the shipment weight was positive, the natural log of weight was compared to the upper and lower bounds for the shipment's SCTG code. The upper and lower bounds for the SCTG code were determined by calculating the natural log of weight for each shipment with reported positive weight data in the SCTG code, calculating the quartiles of the distribution of the transformed weight data, and adding or subtracting three times the interquartile range to the

median. If the shipment's weight was zero or its natural log was outside the bounds, then weight was set for imputation, which was reflected in the R/I code for weight. However, if the shipment's weight was positive and its natural log was within the bounds, then value was set for imputation, which was reflected in the R/I code for value. Based on input from the Census Bureau analysts, we assumed that value was incorrect, unless there was an indication that weight was incorrect.

After the above edit was performed, a second imputation edit was performed to determine if imputation could be performed for a shipment in which its value or weight was missing or zero. For the shipment, if value was missing or zero and weight was positive, an edit was performed on weight, similar to the edit described above, to determine if value could be imputed. The natural log of the shipment weight was compared to the same upper and lower bounds for the shipment's SCTG code. If the shipment's transformed weight was outside the bounds, then value was not set for imputation, and this was reflected in the R/I codes for both value and weight. However, if weight was missing or zero and value was positive, an edit was performed on value, similar to the one just described for weight, to determine if weight could be imputed.

For a given shipment, to perform imputation on value or weight, the other item must be positive and the shipment's SCTG code must be valid. A third imputation edit was performed that identified shipments for which value and weight were both missing or zero, as well as shipments for which the SCTG code was missing or invalid. This was reflected in the R/I codes for both value and weight.

5. Imputation of Shipment Value or Weight Data

Based on the imputation edits described in Section 4.4, we identified shipments, or *recipients*, for which either value or weight was to be imputed. Besides running imputation at the end of a quarter, as described in Section 4.2, we also ran imputation prior to the creation of published estimates. For either value or weight, the item was imputed for about 3% of the 2,649,761 reported shipments, and the item could not be imputed for less than 1% of the shipments.

This section describes the methodology used to impute either value or weight, including the procedure used to match recipients and their potential donors, the selection of a donor for a given recipient, the imputation of the recipient's item using the donor's value and weight data, and the method used when a donor could not be found for a recipient. Because the methodology used to impute a recipient's value is similar to the methodology for weight, we will assume for this discussion that weight is to be imputed.

Nonresponse to both shipment value and weight in the 2002 CFS was addressed by adjusting the sample weights assigned to shipments whose data were included in the estimates. For more information on the nonresponse adjustments used in estimation, see Evans and Cantwell (1995).

For both value and weight, we formed separate classes by SCTG code, based on the percentiles of the reported data by commodity. These classes were used to group donors and recipients having similar value or weight data. We performed an edit, similar to the Value-to-Weight Edit described in Section 4.1, so that shipments, whose value and weight data had been flagged for review and verified by the Census Bureau analysts, would not adversely affect the calculation of the percentiles or become donors. We also formed collapsed classes within a given SCTG code, in the event that a donor could not be found in a recipient's original class.

For a given recipient with weight requiring imputation, we determined a *donor pool* based on shipments that had reported positive value and weight data, were in the recipient's SCTG code, and were similar to the recipient in terms of company affiliation, geographic origin, and value. Within the recipient's value class, we first searched for one or more potential donors from the same establishment, company, or MA code as the recipient. We next searched within the recipient's collapsed value class. Finally, we searched without regard to value.

After the donor pool for a recipient had been determined, the *donor* was randomly selected from the pool. There was no limit to the number of times that a given shipment could be used as a donor. Also, a recipient's donor did not have to be shipped in the same reporting week or quarter as the recipient.

The recipient's weight was imputed by multiplying its value by the weight-to-value ratio from the donor. If no donor could be found for the recipient, its value was multiplied by the median weight-to-value ratio for shipments with reported positive value and weight data in the recipient's SCTG code. When value was imputed from weight, the reciprocal of the median weight-to-value ratio was used if no donor could be found.

Table 3 gives the number of shipments with imputed weight or value by imputation method. Donor imputation within size class was used for approximately 89% of the shipments with imputed weight and 94% of the shipments with imputed value. Donor imputation within collapsed size class was rarely used - only about 2% of the imputed shipments for either weight or value used this method. About 6% of the shipments with imputed weight used a median weight-to-value ratio, compared to about 2% of the shipments with imputed value.

Table 3: Frequencies of Shipments with Imputed Weight or Value by Imputation Method

Imputation Method	Weight	Value
Donor within Size Class	77,660	76,633
Donor within Collapsed Size Class	2,013	1,261
Donor without Regard to Size	2,527	1,453
Median Weight-to-Value Ratio	5,278	2,021
Total	87,478	81,368

6. Research Plans for the 2007 CFS

For the 2007 CFS, we have begun to conduct research in areas related to frame construction, edits, and imputation. This research includes looking into enhancements to the current methods, such as evaluating the industries that define the target population, analyzing the current methods of calculating the MOS, investigating the effectiveness of the current edit techniques, and improving the imputation of value and weight data.

The research also includes looking into new methods. We are investigating the possibility of creating the MOS for mining establishments based on weight data from the 2002 Census. We plan to look into the possibility of replacing the Value-to-Weight Edit with an edit based on methodology proposed by Hidiroglou and Berthelot (1986). We also plan to evaluate the method used to match donors to recipients when imputing value or weight, which may include a match solely on SCTG code as a last resort.

References

Black, J., W. Davie Jr., and R. Detlefsen (2003), "The 2002 Commodity Flow Survey Design Process." *2003 Proceedings of the American Statistical Association*, Section on Survey Research Methods, Alexandria, VA: American Statistical Association: pp. 578-584.

Cochran, W. (1977), *Sampling Techniques.* New York: John Wiley & Sons.

Evans, T. and P. Cantwell (1995), "Adjusting the 1993 Commodity Flow Survey to the 1992 Economic Census." *1995 Proceedings of the American Statistical Association*, Section on Survey Research Methods, Alexandria, VA: American Statistical Association, pp. 896-901.

Hidiroglou, M.A. and J.-M. Berthelot (1986), "Statistical Editing and Imputation for Periodic Business Surveys." *Survey Methodology, June 1986.* Vol. 12, No. 1, pp. 73-83. Statistics Canada.

Mosteller, F. and J.W. Tukey (1977). *Data Analysis and Regression*. Reading, MA: Addison-Wesley Publishing Co., Inc.

U.S. General Accounting Office (2004), "Metropolitan Statistical Areas: New Standards and Their Impact on Selected Federal Programs." Report to the Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census, Committee on Government Reform, House of Representatives.

U.S. Office of Management and Budget (1998). *North American Industry Classification System: United States, 1997*. Lanham, MD: Bernan Press.

U.S. Office of Management and Budget (1987). *Standard Industrial Classification Manual 1987*. Springfield, VA: National Technical Information Service.