THE 2002 COMMODITY FLOW SURVEY DESIGN PROCESS[1]

Jock Black, William C. Davie Jr., Ruth E. Detlefsen, U.S. Bureau of the Census
Jock Black, U.S. Bureau of the Census, SSSD, FOB 2754-3, Washington D.C. 20233

Keywords: Stratification, Sample Design, Establishment Survey

Abstract

The Commodity Flow Survey (CFS), jointly sponsored by the U.S. Census Bureau and the Department of Transportation's Bureau of Transportation Statistics (BTS), is conducted about once every five years to provide data on the flow of commodities by various transportation modes. For the 2002 survey, several different sample design options were considered to address funding uncertainties. The goal was to design a sample the size of the 1997 survey (100,000 establishments) or fewer that would enable us to publish estimates of shipping volumes by geographic areas, transportation modes, and commodities similar to those used in the previous surveys' publications. Toward that end, we investigated the feasibility of a sample of 10,000 establishments, 30,000 establishments, and 50,000 establishments. We evaluated whether it was better to explicitly stratify by metropolitan area or to select a systematic sample from a frame of establishments sorted by metropolitan area. We augmented the sample to increase the likelihood that a smaller sample would still capture shipments for the less prominent modes of transportation. In doing this research we strove to evaluate the effects of proposed design changes on the number and quality of published estimates. This paper will detail research conducted to arrive at the final sample design and describe that design.

Introduction

The Commodity Flow Survey started in 1993 as part of the Census Bureau's Economic Census. The survey coverage was expanded from the former Commodity Transportation Survey's manufacturing coverage to include wholesale, mining, selected retail kinds of business, and warehouses. The 1993 CFS sample was large, consisting of more than 200,000 establishments selected from a universe of over 800,000 establishments. The 1997 CFS sample size was 100,000 establishments - cut back to avoid processing backlogs and allow more thorough editing and review. When planning began for the 2002 survey, BTS was unsure of the CFS funding and asked the Census Bureau to evaluate the potential of samples of 10,000, 30,000 and 50,000 establishments.

The CFS is a multipurpose survey. It attempts to measure shipping dollar volumes, tonnages, and total ton-miles by commodity, shipment origin, shipment destination, and transportation mode with breakouts according to shipment weight and overall shipment distance. Some published tables summarize shipments by four dimensions. The 1997 CFS published more than two thousand tables of estimates! Many of the criteria used for classifying shipments for the 1997 tables are not available from administrative sources. The sample design and estimation are complex. Basic sampling formulae to evaluate the decrease in sample size would not have been appropriate. Rather, we opted to simulate the results of the different sample sizes using 1997 data and evaluated how the publications would likely change.

Overview of the CFS Sample Design and Estimation

The sample for the 1997 Commodity Flow Survey (CFS) was selected using a stratified three-stage design in which the first-stage sampling units were establishments, the second-stage sampling units were 1-week periods (reporting weeks) within the 13 weeks of each quarter of 1997, and the third-stage sampling units were shipments. First-stage stratification variables included industry (based on the Standard Industrial Code (SIC)), size (estimated value of shipments), and geography (based on National Transportation Analysis Regions (NTARs)). We planned to use a similar design for the 2002 survey although SIC had been superseded by the North American Industry Classification System (NAICS) and NTARS were no longer relevant. A thorough description of the final 2002 CFS design will be presented later in this paper.

Briefly, estimates for the 1997 survey are based on individually weighted shipments. Each shipment's weight is the product of the reciprocals of the probabilities of selection for each of the three stages of sampling, nonresponse adjustment factors for the last two stages, and two census-adjustment factors. One of the census adjustment factors ensures that the weighted dollar value of shipments from an establishment is identical to the establishment's total value of shipments derived from the 1997 Economic Census. The other

[1]This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

census adjustment factor ensures that the weighted value of shipments for all establishments in a particular kind of business is identical to the total value of shipments for the same kind of business derived from the 1997 Economic Census. For details, see Evans, 1995. The census adjustments make it particularly difficult to predict how changes in sample size will affect the variances of the final estimates.

Simulation Description - Sample Size

The plan for evaluating the effects of the suggested sample sizes was fairly simple. Select a sample of establishments from the 1997 survey and use the data reported from those establishments to retabulate a variety of tables published in 1997. We would keep track of what was originally published and what would now be published. This approach would allow us to realize the benefits and drawbacks of a real survey, e.g., nonresponse, editing, imputation, and census adjustment. It is important to note that the Census Bureau suppresses publication of estimates that have large imputation rates or large sampling variances. Estimates that risk disclosing information that would allow a particular respondent to be identified are also suppressed. For the CFS, the reduction in sample size would primarily result in increases in sampling variance and thus fewer publishable estimates.

The difficulty in carrying out the plan was in designing a subsample that would reasonably mimic a realistic design. Weights for the subsample should reasonably approximate weights from a stratified random sample design. Reducing the sample size would dramatically reduce stratification detail. Ultimately, we decided that the subsample would use a stratified, systematic, probability proportional to estimated size design. The measure of size used to select establishments was the reciprocal of the establishment's first-stage selection probability from the 1997 CFS.

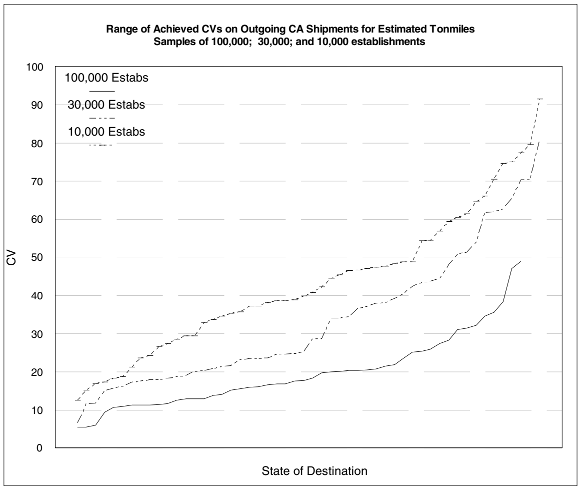
The table in the adjacent column compares the 1997 CFS design characteristics with those used for the hypothetical design used to evaluate the 30,000 establishments scenario.

Similar decisions were implemented to evaluate the 10,000 and 50,000 establishment scenarios.

Characteristic	1997 CFS	Scenario one
Sample Size	100,000	30,000
Certainty Determination	Measure of size cutoffs were determined at 3-digit SIC level 57,000 Certainties	Largest 0.5% of establish-ments in each state and any 1997 certainty warehouses 3,000 Certainties + 7,000 PPS certainties
Noncertainty Stratification - Primary Strata	44 industry strata based on 3-digit SIC in mining, manu-facturing, wholesale, and retail 89 NTARs Auxiliaries stratified by trade area 4,000 Primary Strata	6 industry strata based on first digit of SIC code State (+ DC) Auxiliaries stratified only by type of auxiliary (Warehouse, Central Admin-istrative Office, other) 180 Primary Strata
Noncertainty Stratification - Secondary Strata and Allocation of Sample	Dalenius and Hodges Rule Maximum of 6 secondary size strata Allocation to estimate U.S. Total Value of Shipments	Same as 1997
Sample Selection	Simple Random Sampling within Strata	Probability Proportional to 1997 CFS weight

Simulation to Determine Sample Size

We found, as expected, that the sampling variability (as measured by the coefficient of variation) increased with decreases in sample size. The following chart gives an example of the increases that could be expected for samples of 10,000 and 30,000 establishments.



In summary, under the 10,000-establishment scenario, the only estimates that could be published with coefficients of variation (CVs) less than 50% were for flows (estimates based on both shipment origin and destination) at the divisional or regional level. The CVs for most mode and commodity estimates would likely be greater than 50%. CVs for broader characterizations of the data such as total value of shipments interstate vs. intrastate, may be lower.

Under the 30,000-establishment scenario, state-to-state flows would likely have CVs less than 50% for most states. The accuracy of mode and commodity detail estimates was less certain. To have estimates with reliability less than 50% would likely require collapsing stub detail. Truck and rail modes were generally well measured but detailed estimates by distance or weight might have high CVs. Most estimates for water modes, alone or in combination, as well as other less frequently used modes, would have CVs greater than 50%.

A sample of 50,000 establishments would give us at the 1) national level, data comparable to the 1997 sample with decreased reliability (CVs about 10% greater); 2) regional or divisional level, data comparable to the 1997 sample with less reliability (CVs about 15% greater); 3) state level, basic one-dimensional tables comparable to the 1997 sample with less reliability (CVs 15% greater). Tables in two or more dimensions would be of questionable reliability. Metropolitan level tables comparable to 100,000 sample with less reliability (CVs 20% greater).

Simulation 2 - Number of Quarters

Ultimately, an agreement was reached to fund a sample of 50,000 establishments canvassed quarterly - a total of 200,000 collections. However, during the survey planning phase we evaluated whether we could expect more reliable estimates by canvassing 100,000 establishments two times a year.

It can be argued that one might expect better results using the second method. For a three-stage simple random sample without replacement design (SRSWOR), we have:

$$V(\bar{\bar{\bar{y}}}) = \left(\frac{N-n}{N}\right)\frac{S_1^2}{n} + \left(\frac{M-m}{M}\right)\frac{S_2^2}{mn} + \left(\frac{P-p}{P}\right)\frac{S_3^2}{pmn}.$$

Where N = first stage universe count, n = first stage sample size, M = second stage universe count for each first stage unit, m = second stage sample size for each first stage unit, P = third stage universe count for each second stage unit, i.e., assume same number of shipments for each week for each establishment, p = third stage sample size for each second stage unit.

For the CFS, N would be establishments, M would be weeks in 2002, and P would be the shipments in a given week for a particular establishment. The CFS design is not a three-stage SRSWOR, but this equation can be used to get some idea of how to allocate the first and second stage units. The second stage of sampling in the CFS is a systematic sample of one from the collection of the thirteen week-in-quarter periods available during 2002. However, the SRSWOR variance formula is probably a reasonable approximation to the systematic variance. The formula above may overstate the variance of a systematic sample. The actual third stage in CFS doesn't satisfy the requirement that there are the same number of shipments for each week for each establishment. Nevertheless, the contribution of the third stage of sampling to the variance of an estimate will be some value that is unchanged as long as mn, the number of establishment quarters measured, is a constant.

Allowing these simplifications, to determine how to allocate a given number of establishment quarters to establishments and quarters, we need to find the minimum of the equation with respect to m and n, given that m*n=k, and m and n are positive integers. Note that the third stage variance depends only on mn, and is unchanged as long as k is unchanged.

Substituting m=k/n, into the above and simplifying, we get

$$V(\bar{\bar{\bar{y}}}) = \left(\frac{1}{n}\right)\left(S_1^2 - \frac{S_2^2}{M}\right) - \left(\frac{S_1^2}{N} - \frac{S_2^2}{k} - V_{3rd\ stage}\right).$$

The second term on the right is unchanged as long as k is held constant. Thus the variance of the estimated mean (and total) decreases with increasing n if we assume that the term in parentheses is positive. We conclude the minimum variance is achieved by setting m=1 and n=k. In our case, 200,000 establishments sampled once should have the smallest variance compared to the 100,000 x 2, or 50,000 x 4 options.

However, because we had already done several simulations, we investigated this proposal using the data collected from the 1997 survey. It used a 100,000 x 4 scheme. For each establishment we randomly selected its reports for either the first and third quarter or the second and fourth quarter and retabulated the estimates. Interestingly, we found that the 50,000 x 4 design actually resulted in more estimates with lower CVs. In addition, we strongly felt that nonsampling error would be much less with four measurements per establishment because of the opportunity to follow-up and correct questionable reports. This table shows the median CVs for three types of estimates (Value (V), Tonnage (T) and Ton-miles (TM)) and the percent of flows having CVs less than 50% for three designs for a table that measures shipments by origin and destination state.

Design	Est	Median CV	Flows Measured with a CV < 50%
100,000 x 4	V	17.8	95%
50,000 x 4	V	19.9	94%
100,000 x 2	V	21.7	89%
100,000 x 4	T	22.6	89%
50,000 x 4	T	26.7	85%
100,000 x 2	T	29.1	81%
100,000 x 4	TM	22.9	88%
50,000 x 4	TM	27.0	86%
100,000 x 2	TM	29.5	80%

In summary, the final decision was to implement a sample of 50,000 establishments that would be canvassed once each quarter of 2002. The remainder of this paper summarizes the creation of the sampling frame and the design and selection of the sample for each stage.

It should be noted that due to the limited time in which to do the evaluations and the complexity of the design and issues involved, the simulations were not based on multiple subsamples. If we had more time, a more rigorous approach would have been taken. We could have investigated various stratification and sampling schemes and performed multiple simulations.

2002 CFS Design - First Stage - Sampling Frame

To create the first-stage sampling frame, we extracted a subset of establishment records from the Business Register as of September 2001. Establishments located in the United States, having nonzero payroll in 2000, and classified in mining (except oil and gas extraction), manufacturing, wholesale, or electronic shopping and mail order retail sectors, as defined by the 1997 North American Industry Classification System (NAICS), were included on the sampling frame. *Auxiliary establishments* (e.g., warehouses and central administrative offices) with shipping activity were also included on the sampling frame. Auxiliary establishments are establishments that are primarily involved in rendering support services for other establishments within the same company, instead of for the public, government, or other business firms. All other establishments included on the sampling frame are referred to as *nonauxiliary establishments*.

Because of changes between the 1997 NAICS and the 2002 NAICS, an unknown portion of establishments classified in the Retail Trade sector in the 1997 Economic Census are expected to be classified in the Wholesale Trade sector in the 2002 Economic Census. Because we wanted complete coverage of the Wholesale Trade sector (as defined for the 2002 Economic Census), the sampling frame also included retail establishments with characteristics likely to indicate reclassification to wholesale for the 2002 Census.

Establishments classified in agriculture, forestry, fishing, utilities, construction, transportation, services, and all other retail industries were not included on the sampling frame. The resulting frame comprised 757,467 establishments, each classified in one of 575 six-digit NAICS codes.

For each establishment we extracted census sales, payroll, and number of employees, name and address, a 6-digit NAICS code, and a primary identifier. Using census sales, payroll, and some additional data collected in the 1997 Economic Census, a measure of size was computed for each establishment. The measure of size was designed to approximate an establishment's annual total value of shipments.

All of the establishments included on the sampling frame had state, county, and place geographic codes.

We used these codes to assign each establishment to one of the 273 metropolitan areas (MAs) defined as a combination of the metropolitan statistical areas (MSAs) and consolidated metropolitan statistical areas (CMSAs).

For the sample design process, we used an estimate of annual total value of shipments as the measure of shipping volume for each establishment. The geographic stratification was based on a combination of the 50 States and the District of Columbia and the 50 most populous metropolitan areas (based on population in the 2002 Decennial Census). As a proxy for commodity when designing the sample, we used combinations of 3- and 4-digit NAICS codes assigned to each establishment.

General Design Research and Decisions

As a guide, we wanted to select an average of 25 - 35 establishments per primary stratum so that the Central Limit Theorem applies to the distribution of our estimates and inferences from these estimates are more robust. With a total sample size of approximately 50,000 establishments, we would have about 1,600 - 2,000 primary strata. Primary strata would be defined by geography and industry, industry serving as a proxy for commodity. The minimal geography would be the 50 States and the District of Columbia. In this instance, the number of industry strata must be in the range 32 to 40 to meet the desired number of primary strata. Incorporating 50 metropolitan areas (MAs) into this geographic stratification would limit the number of industry strata to a range of 16 to 20.

Following this reasoning, we considered three alternative sample designs to determine which best measured the value of shipments for MAs. The essential difference between the designs was in how establishments located in the top 50 MAs would be sampled. For the first design, primary strata were defined by the 50 States and DC, the top 50 MAs, and NAICS industry recodes. Establishments within each primary stratum were stratified by measure of size (or substratified) to improve the efficiency of the sample. Then, we selected a simple random sample of establishments without replacement from each substratum. For the second and third designs, we excluded metropolitan area from the primary stratification. Thus, primary strata were defined by the 50 States and DC and NAICS industry recodes. Because we used only 51 geographic strata in this design, the industry stratification was much more detailed. As in the first design, we substratified the establishments within each primary stratum by measure of size. The second design selected a simple random sample (SRS) without replacement from each substratum. For the third design, we selected a *systematic* sample (SYS) of establishments from each substratum. Prior to selection, establishments in each substratum were sorted by a variable to indicate if the establishment was located in one of the top 50 MAs. By sorting and performing a systematic selection we hoped to achieve reasonable coverage of the MAs.

We selected several samples using each of these designs. Based on a comparison of the coefficients of variation (CVs) achieved for estimated total value of shipments (the estimated total measure of size) at 3- and 4-digit NAICS levels and at state and MA levels, we selected the first design. For the SRS and the SYS design with state x industry stratification the average CV on estimated total measure of size across all MAs was 3.7%, across all industries it was 1.0%. For the SRS design with the state x MA x industry stratification those numbers were 0.9% and 2.1%, respectively.

We felt the improved precision of the MA estimates for this design outweighed the slight loss in precision of the industry level estimates. Furthermore, the precision of the state-level estimates was similar between the two designs (average 0.6% CV). A complete description of the design we chose is provided next.

Primary Stratification

A total of 110 geographic strata was defined by the intersections of the 50 States and the District of Columbia with the top 50 MAs. Several MAs cross state boundaries and were responsible for more than one state x MA stratum. If a particular MA was not one of the 50 largest, then it was collapsed with the remaining MAs and non-MAs within its state. We referred to these collapsed strata as "Rest of State (ROS)" strata. When a MA crossed state boundaries, we considered the size of each part of the MA relative to the MA's total measure of size when determining whether or not to create strata in each state in which the MA was defined. Industry strata were created as follows. Within each of the 110 geographic strata, we started with a total of 45 industry groups: 3 mining (4-digit NAICS); 21 manufacturing (3-digit NAICS); 18 wholesale (4-digit NAICS); 1 retail (NAICS 4541); and 2 auxiliary (NAICS 4931 and 5511). We then implemented a rule that states a particular industry stratum will be defined within a geographic stratum if it contributes at least 2 percent to its corresponding State total measure of size or it contributes at least 2 percent to the national total measure of size for the industry. Industry groups not meeting these criteria were combined into at most 12 new "collapsed" industry strata using a hierarchical clustering algorithm.

We wanted to group NAICS codes in such a way that we would have stratification by commodity. To do this, the clustering algorithm required measures of the volume of each commodity being shipped by establishments in each NAICS industry. This type of information was available to us from the 1997 CFS. However, the 1997 CFS used the SIC system to classify establishments and we required NAICS codes for the 2002 CFS. We attempted to obtain a NAICS code for each establishment tabulated in the 1997 CFS by matching a file of these establishments to the 2002 CFS sampling frame. From this matching operation, we were able to retrieve a NAICS code for approximately 66% of the tabulated establishments. Using all tabulated shipments associated with the matched establishments, we created total value and tons estimates by 2-digit commodity for each of the 45 NAICS industry groups previously identified. For each industry, we transformed the totals to percentages by computing the percent of total value and percent of total tons contributing to each commodity. Converting the totals to percentages normalized the representation of each commodity in the clustering algorithm. Had we not done this, the most expensive or heavy commodities would have controlled the clustering. These percentages were used as input to the clustering algorithm.

Because of potential differences in shipping patterns between auxiliary and nonauxiliary establishments, we created two industry strata of auxiliary establishments in every geographic stratum. The final design used 2,332 primary strata.

Certainty Decisions

Within each primary stratum, we used the Lavalier-Hidiroglou (L-H) algorithm to determine the lower boundary of each certainty stratum. The parameters required by the program were set so that about 20,000 establishments (or 40% of the total sample size) would be selected with certainty. The decision to select about 40% of the sample with certainty was a compromise between the percent of certainties selected in prior CFS samples and the results of the L-H algorithm using the same CV constraint for each stratum. To achieve the desired number of certainties, we adjusted the CV constraint input to the L-H algorithm for selected groups of states. At this step, we selected about 23,200 establishments with certainty compared to just over 57,000 establishments with certainty (about 56% of the total sample size) for the 1997 CFS sample.

Because the 2002 sample was about half the size of the 1997 CFS sample, we were concerned about the ability of the sample to capture less frequent types of shipments (e.g., air, water, rail, and hazardous materials). We considered over-sampling in some strata, but based on research conducted using shipment data collected in the 1997 CFS, we did not find any industries which had a substantially greater relative proportion of such shipments. Ultimately, we felt the best approach was to identify those establishments which made the bulk of these types of shipments in 1997 and then select them with certainty.

To identify these establishments, we proceeded as follows. We identified all establishments in the 1997 CFS sample that reported shipments made by air, water, or rail. We also identified those establishments that reported hazardous materials shipments. For each of these establishments, we computed the contribution to the establishment's total value and tonnage accounted for by each of these types of shipments. Next, we matched these establishments to the final sampling frame for the 2002 CFS and assigned each establishment to one of three groups: (1) MOS greater than the certainty boundary, (2) MOS less than the certainty boundary, or (3) not on the sampling frame. We then looked to see what percent of the total volume (value or tons) of shipments for each type of shipment was captured by selecting with certainty the largest 50, largest 100, or all establishments in the second group. Largest was defined in terms of value and tons of each type of shipment (air, water, rail, or hazardous). There was little benefit to adding more than the 100 largest shippers (by value and tonnage) for each category. Once these 800 establishments were identified, we grouped them into one file and unduplicated them. This procedure added a total of about 500 certainty establishments.

Noncertainty Decisions

The remaining establishments made up the noncertainty sampling frame. Within each primary stratum, we used the Dalenius-Hodges cumulative $\sqrt{f}$ rule to set boundaries for 1 to 6 measure-of-size strata, or substrata. We then used Neyman allocation to determine the sample size required within each substratum to meet a CV constraint on the estimated primary stratum total. The final design used a total of 9,577 substrata. For each noncertainty substratum, we set three constraints: (1) a minimum sample size of two establishments; (2) a maximum sampling weight of 100 (minimum sampling fraction of 1/100); (3) and the stratum sample sizes had to be integers. The second constraint was implemented to rein in the possibility of the third stage of sampling resulting in very large final shipment weights. These steps were repeated several times, varying the coefficient of variation constraint each time, until a total sample size of about 50,000 establishments was achieved. This was achieved with a CV of 4.9 % for each primary stratum.

Sample Selection
Within each substratum, we selected a simple random sample of establishments without replacement. The final sample contained 50,956 establishments.

Proof of Sample
As part of our review of the selected sample of establishments, we examined several statistics. We calculated Z-statistics to measure the standardized distance between the estimated total measure of size (MOS), computed from the selected sample, and the total measure of size, computed from all establishments included on the sampling frame. Each Z-statistic was computed as

$$Z = \frac{m\hat{o}s - mos}{SE(m\hat{o}s)}.$$

Z-statistics were calculated at the primary stratum, state, metropolitan area, and 3-, 4-, and 6-digit NAICS levels. We also looked at the coefficient of variation of the estimated total MOS and the sampling rates at each of these levels. Finally, we reviewed the ratio of the estimated total MOS to the total MOS computed from all establishments on the sampling frame at these same levels. These ratios should be close to one. After the review, we felt that the selected sample met the objectives of the survey.

Second-Stage
The frame for the second stage of sampling consisted of 13 one-week periods (reporting weeks) within each quarter of 2002. Each establishment selected into the 2002 CFS sample was systematically assigned to report for four reporting weeks - one in each quarter of the reference year. Each of the four weeks was in the same relative position of the quarter. For example, an establishment might have been requested to report data for the 5^{th}, 18^{th}, 31^{st}, and 44^{th} weeks of the reference year. In this instance, each reporting week corresponds to the 5^{th} week of each quarter. Prior to assignment of weeks to establishments, we sorted the selected sample by primary stratum (state x metropolitan area x industry) and measure-of-size stratum (substratum).

Third Stage
For each of the four reporting weeks in which an establishment was asked to report, we requested the respondent to construct a sampling frame consisting of all shipments made by the establishment in the reporting week. Each respondent was asked to count the total number of shipments comprising the sampling frame and to record this number on the questionnaire. For each assigned reporting week, if an establishment made *more than 40* shipments during that week, we asked the respondent to select a systematic sample of these shipments and to provide us with information only about the selected shipments. If an establishment made *40 or fewer* shipments during that week, we asked the respondent to provide information about *all* of the establishment's shipments made during that week, i.e., no sampling was required.

The respondent used a table provided to determine the sampling rate or take-every interval for sampling shipments. The size of a particular respondent's sample for a given reporting week was to be between 20 and 40 shipments, inclusive, depending on the total number of shipments the establishment made during a reporting week. To select the sample, we instructed the respondent to count the shipments in the sampling frame, starting with the first. When the respondent reached the selection rate, that shipment was selected into the sample. The respondent was to continue counting with the next shipment, counting it as one. When the selection rate was reached again, that shipment was selected into the sample. This procedure was repeated until the last shipment on the sampling frame was counted.

References
Evans, B. Timothy & Patrick J Cantwell. (1995) "Adjusting the1993 Commodity Flow Survey to the 1992 Economic Census." *American Statistical Association 1995 Proceedings of the Survey Research Methods Section.* Alexandria, VA: American Statistical Association, pp. 216-221.

Smith, Robert T., Evans, B. Timothy, & Fowler, John L. (1994) "The Commodity Flow Survey." *American Statistical Association 1994 Proceedings of the Business and Economic Statistics Section.* Alexandria, VA: American Statistical Association, pp. 89-94.