2007 Commodity Flow Survey Overview and Methodology
United States. Census Bureau

from: https://www.census.gov/programs-surveys/cfs/technical-documentation/methodology/methodology-2007.html
[Last Revised: January 25, 2017]

National Transportation Library preservation note: The information in this document was originally presented on the United States Census Bureau website, and was last accessed on August 31, 2021, at the URL: https://www.census.gov/programs-surveys/cfs/technical-documentation/methodology/methodology-2007.html

Because this information is important context for the 2007 Commodity Flow Survey, funded in part by the Bureau of Transportation Statistics (BTS), NTL is providing access to the methodology information for long-term preservation purposed.

Please note that while the information gathered below is not displayed in the same way as the original webpages, the intellectual content of those pages is captured and preserved for future use. Minor edits were introduced to improve navigation or readability.

The actually Publication Date is not known, and we have estimated a publication date of January 25, 2017, from the webpage update note at the bottom of the page.

These preservation actions were taken on August 31, 2021.

2007 Commodity Flow Survey Overview and Methodology
United States. Census Bureau

from: https://www.census.gov/programs-surveys/cfs/technical-documentation/methodology/methodology-2007.html

In This Section:

Overview

The Commodity Flow Survey (CFS) is a joint effort by the Research and Innovative Technology Administration (RITA), Bureau of Transportation Statistics (BTS), and the U.S. Census Bureau, U.S. Department of Commerce. The first CFS was conducted in 1993, followed by surveys in 1997, 2002, and 2007. Since 1997, the survey has been conducted in years ending in "2" or "7," aligning it with (and as a component of) the Economic Census. The survey produces data on the movement of goods in the United States. It provides information on the type, value, weight, origin and destination, and mode of transportation of commodity shipments originating from manufacturing, mining, wholesale, and select retail establishments located in the 50 states and the District of Columbia. The CFS data are used by policy makers and transportation planners in various federal, state, and local agencies for assessing the demand for transportation facilities and services, energy use, and safety risk and environmental concerns. Additionally, business owners, private researchers, and analysts use the CFS data for analyzing trends in the movement of goods, mapping spatial patterns of commodity and vehicle flows, forecasting demands for the movement of goods, and determining needs for associated infrastructure and equipment.

Back to top

Objective

The primary objective for the 2007 Commodity Flow Survey (CFS) was to estimate shipping volumes (value, tons, and ton-miles) by commodity and mode of transportation at varying levels of geographic detail. A secondary objective was to estimate the volume of shipments moving from one geographic area to another (i.e., flows of commodities between states, regions, etc.) by mode and commodity. A detailed description of the survey coverage and sample design for the 2007 CFS is provided below.

Back to top

Data Collection

Each establishment selected into the CFS sample was mailed a questionnaire for each of its four reporting weeks, that is, an establishment was sent a questionnaire once every quarter of 2007. For a given establishment, the respondent was asked to provide the following information about each of the establishment's reported shipments:

- Shipment ID number
- Shipment date (Month, Day)
- Shipment value
- Shipment weight in pounds
- Commodity code from Standard Classification of Transported Goods (SCTG) list
- Commodity description
- United Nations or North America (UN/NA) number for hazardous material shipments
- U.S. destination (city, state, zip code) - or gateway for export shipment
- Modes of transport
- An indication of whether the shipment was an export
- City and country of destination for exports
- Export mode

For a shipment that included more than one commodity, the respondent was instructed to report the commodity that made up the greatest percentage of the shipment's weight.

Back to top

Imputation of Shipment Value or Weight

Only two items were ever imputed in the 2007 CFS - shipment value or weight. To correct for nonresponse to either the value or weight for a given shipment reported in the CFS, the missing value for the item (or value that failed edit) was replaced by a predicted value obtained from an appropriate model. Such a shipment was considered a "recipient" if it had a valid commodity code and the other item reported was greater than zero and had passed edit. The recipient's item that was missing or failed edit was imputed as follows. First, a "donor" shipment was randomly selected from shipments that were reported in the CFS with:

- The same commodity code as the recipient.
- Both value and weight items reported greater than zero and had passed edit.
- Similar origin and value for the item reported by the recipient.

Then, the donor's value and weight data were used to calculate a ratio, which was then applied to the recipient's reported item, to impute the item that was missing or failed edit. If no donor was found, the median ratio for all shipments reported in the survey with the same commodity code as the recipient - and with both value and weight items reported greater than zero - was applied to the recipient's reported item. For either the value or weight item, about three percent of the shipment records used for the calculation of estimates had imputed data for the item.

Back to top

Estimation

Estimated totals (e.g., value of shipments, tons, ton-miles) were produced as the sum of weighted shipment data (reported or imputed). Percent change and percent-of-total estimates were derived using the appropriate estimated totals. Estimates of average miles per shipment were computed by dividing an estimate of the total miles traveled by the estimated number of shipments.

Each shipment had associated with it a single tabulation weight, which was used in computing all estimates to which the shipment contributes. The tabulation weight was a product of seven different component weights. A description of each component weight follows.

CFS respondents provided data for a sample of shipments made by their respective establishments in the survey year. For each establishment, an estimate of that establishment's total value of shipments was produced for the entire survey year. To do this, four different weights were used - the shipment weight, the shipment nonresponse weight, the quarter weight, and the quarter nonresponse weight. Three additional weights were then applied to produce estimates representative of the entire universe - the establishment-level adjustment weight, the establishment (or sample) weight, and the industry-level adjustment weight.

Like establishments, shipments were identified as either certainty or noncertainty (see the Nonsampling Error section below). For noncertainty shipments, the **shipment weight** was defined as the ratio of the reported total number of shipments made by an establishment in a reporting week to the number of sampled shipments for the same week. This weight used data from the sampled shipments to represent all the establishment's shipments made in the reporting week. However, a respondent may have failed to provide sufficient information about a particular sampled shipment. For example, a respondent may not have been able to provide value, weight, or a destination for one of the sampled shipments. If this data item could not be imputed, then this shipment did not contribute to tabulations and was deemed unusable. (A *usable shipment* is one that has valid entries for value, weight, and origin and destination ZIP Codes.) To account for these unusable shipments, a **shipment nonresponse weight** was applied. For noncertainty shipments from a particular establishment's reporting week, the weight was equal to the ratio of the number of sampled shipments for the reporting week to the number of usable shipments for the same week. The shipment weight for certainty shipments from a particular establishment's reporting week was equal to one.

The **quarter weight** inflated an establishment's estimate for a particular reporting week to an estimate for the corresponding quarter. For noncertainty shipments, the quarter weight was equal to 13. The quarter weight for most certainty shipments is also equal to 13. However, if a respondent was able to provide information about all large (or certainty) shipments made in the quarter containing the reporting week, then the quarter weight for each of these shipments was one. For each establishment, the quarterly estimates were added to produce an estimate of the establishment's value of shipments for the entire survey year. Whenever an establishment did not provide the Census Bureau with a response for each of its four reporting weeks, a quarter nonresponse weight was computed. The **quarter nonresponse weight** for a particular establishment was defined as the ratio of the number of quarters for which the establishment was in business in the survey year to the total number of quarters (reporting weeks), for which usable shipment data was received from the establishment.

Using these four component weights, an estimate of each establishment's value of shipments was computed for the entire survey year. This estimate was then multiplied by a factor that adjusts the estimate using value of shipments and sales data obtained from other surveys and censuses conducted by the Census Bureau. This weight, the **establishment- level adjustment weight**, attempted to correct for any sampling or nonsampling errors that occurred during the sampling of shipments by the respondent.

The adjusted value of shipments estimate for an establishment was then weighted by the **establishment (or sample) weight**. This weight was equal to the reciprocal of the establishment's probability of being selected into the first stage sample.

A final adjustment weight, the **industry- level adjustment weight**, used information from other surveys and censuses conducted by the Census Bureau to account for establishment nonresponse or non-useable response, and for changes in the universe of establishments from 2006 when the first-stage sampling frame was constructed and 2007 the year in which the data were collected. Separate industry-level adjustment weights were determined for nonauxiliary and auxiliary establishments.

Back to top

Reliability of the Estimates

The estimates presented by the 2007 CFS may differ from the actual, unknown population values. Statisticians define this difference as the total error of the estimate. When describing the accuracy of survey results, it is convenient to discuss total error as the sum of sampling error and nonsampling error. Sampling error is the average difference between the estimate and the result that would be obtained from a complete enumeration of the sampling frame conducted under the same survey conditions. Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate.

The sampling error of the estimates in this publication can be estimated from the selected sample because the sample was selected using probability sampling. Common measures related to sampling error are the sampling variance, the standard error, and the coefficient of variation (CV). The sampling variance is the squared difference, averaged over all possible samples of the same size and design, between the estimator and its average value. The standard error is the square root of the sampling variance. The CV expresses the standard error as a percentage of the estimate to which it refers.

Nonsampling errors are difficult to measure and can be introduced through inadequacies in the questionnaire, nonresponse, inaccurate reporting by respondents, errors in the application of survey procedures, incorrect recording of answers, and errors in data entry and processing. In conducting the 2007 CFS, every effort was made to minimize the effect of nonsampling errors on the estimates. Data users should take into account both the measures of sampling error and the potential effects of nonsampling error when using these estimates.

More detailed descriptions of sampling and nonsampling errors for the 2007 CFS are provided in the following sections.

Back to top

Sampling Error

Because the estimates are based on a sample, exact agreement with results that would be obtained from a complete enumeration of all shipments made in 2007 from all establishments included on the sampling frame using the same enumeration procedures is not expected. However, because probability sampling was used at each stage of selection, it is possible to estimate the sampling variability of the survey estimates. For CFS estimates, sampling variability arises from each of the three stages of sampling.

The particular sample used in this survey is one of a large number of samples of the same size that could have been selected using the same design. If all possible samples had been surveyed under the same conditions, an estimate of a population parameter of interest could have been obtained from each sample. These samples give rise to a distribution of estimates for the population parameter. A statistical measure of the variability among these estimates is the *standard error*, which can be approximated from any one sample. The standard error is defined as the square root of the variance. The *coefficient of variation* (CV, or relative standard error) of an estimator is the standard error of the estimator divided by the estimator. For the 2007 CFS, the coefficient of variation also incorporates the effect of the noise infusion disclosure avoidance method. Note that measures of sampling variability, such as the standard error and coefficient of variation, are estimated from the sample and are also subject to sampling variability, and technically they should have been referred to as estimated standard error and estimated coefficient of variation. However, for the sake of brevity, we have omitted this detail. It is important to note that the standard error only measures sampling variability. It does not measure systematic biases of the sample. Individuals using estimates contained in this report are advised to incorporate this information into their analyses, as sampling error could affect the conclusions drawn from these estimates.

An estimate from a particular sample and the standard error associated with the estimate can be used to construct a confidence interval. A confidence interval is a range about a given estimator that has a specified probability of containing the result of a complete enumeration of the sampling frame conducted under the same survey conditions. Associated with each interval is a percentage of confidence, which is interpreted as follows. If, for each possible sample, an estimate of a population parameter and its approximate standard error were obtained, then:

1. For approximately 90 percent of the possible samples, the interval from 1.645 standard errors below to 1.645 standard errors above the estimate would include the result as obtained from a complete enumeration of the sampling frame conducted under the same survey conditions.

2. For approximately 95 percent of the possible samples, the interval from 1.96 standard errors below to 1.96 standard errors above the estimate would include the result as

obtained from a complete enumeration of the sampling frame conducted under the same survey conditions.

To illustrate the computation of a confidence interval for an estimate of total value of shipments, assume that an estimate of total value is $10,750 million and the coefficient of variation for this estimate is 1.8 percent, or 0.018. First obtain the standard error of the estimate by multiplying the value of shipments estimate by its coefficient of variation. For this example, multiply $10,750 million by 0.018. This yields a standard error of $193.5 million. The upper and lower bounds of the 90-percent confidence interval are computed as $10,750 million plus or minus 1.645 times $193.5 million. Consequently, the 90-percent confidence interval is $10,432 million to $11,068 million. If corresponding confidence intervals were constructed for all possible samples of the same size and design, approximately 9 out of 10 (90 percent) of these intervals would contain the result obtained from a complete enumeration.

Back to top

Nonsampling Error

Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate and may also occur in censuses. It is often helpful to think of nonsampling error as arising from deficiencies or mistakes in the survey process. In the CFS, nonsampling error can be attributed to many sources: inability to obtain information about all units in the sample; response errors; differences in the interpretation of the questions; mistakes in coding or keying the data obtained; and other errors of collection, response, coverage, and processing. Although no direct measurement of the potential biases due to nonsampling error has been obtained, precautionary steps were taken in all phases of the collection, processing, and tabulation of the data in an effort to minimize their influence. Individuals using estimates in this report should incorporate this information into their analyses, as nonsampling error could affect the conclusions drawn from these estimates.

A potential source of bias in the estimates is nonresponse. Nonresponse is defined as the inability to obtain all the intended measurements or responses from all units in the sample. Four levels of nonresponse can occur in the CFS: item, shipment, quarter (reporting week), and establishment. Item nonresponse occurs either when a question is unanswered or the response to the question fails computer or analyst edits. Nonresponse to the shipment value or weight items is corrected by imputation, which is the procedure by which a missing value is replaced by a predicted value obtained from an appropriate model.

Shipment, quarter, and establishment nonresponse are used to describe the inability to obtain any of the substantive measurements about a sampled shipment, quarter, or establishment, respectively. Shipment and quarter nonresponse are corrected by reweighting. Reweighting allocates characteristics to the nonrespondents in proportion to the characteristics observed for the respondents. The amount of bias introduced by this nonresponse adjustment procedure depends on the extent to which the nonrespondents differ, characteristically, from the respondents. Establishment nonresponse is corrected during the estimation procedure by the industry-level adjustment weight. In most cases of establishment nonresponse, none of the four questionnaires have been returned to the Census Bureau, after several attempts to elicit a response. Approximately 67 percent of the establishments provided at least one quarter of data that contributed to these tables.

Some possible sources of bias that are attributed to respondent-conducted sampling include misunderstanding the definition of a shipment, constructing an incomplete frame of shipments from which to sample, ordering the shipment sampling frame by selected shipment characteristics, and selecting shipment records by a method other than the one specified in the questionnaire's instructions. The respondents who had reported a shipment with untypically large value or weight when compared to the rest of their reported shipments were often contacted for verification. In such cases, if we were able to collect information on all of the of the large shipments a respondent had made either for a particular reporting week or for the entire quarter, then we identified those large shipments as certainty shipments.

Back to top

Mileage Calculation

The miles traveled by each shipment was determined using the shipment information reported by the respondents, and a software tool, called GeoMiler, that has been developed by the Bureau of Transportation Statistics (BTS) in partnership with MacroSys Research and Technology (MacroSys) for estimating freight travel. GeoMiler calculated miles traveled using Geographic Information System (GIS) technology and spatial multimodal network databases. It integrated map-visualization features with route solvers to handle many alternative multimodal combinations. This tool used algorithms that found the "best path" over spatial representations of the U.S. highway, railway, waterway, and airway networks. For waterborne export shipments, GeoMiler used a waterborne commerce database from the U.S. Army Corps of Engineers to route freight originating in the U.S via the deep sea (ocean). For airborne export shipments, GeoMiler used a newly developed air export network from the RITA/BTS Office of Airline Information (OAI).

For a domestic shipment, the mileage was calculated between the centroid (center of the geographic area) of the U.S. origin ZIP Code and the centroid of the destination ZIP Code. The route between an O-D pair was composed of a series of links, and an impedance factor was assigned to each link (impedance is defined as a function of distance and travel time). Given a mode or modal sequence, the role of GeoMiler was to find that "best path" route which minimized the summed total impedance of the links between the specified O-D pair.

The mileage for shipments within a ZIP Code (matching O-D pair) was calculated by means of a formula that approximated the longest distance within the boundaries of that ZIP Code.

For multimodal shipments (that is, those shipments involving more than one mode, such as truck-rail shipments), spatial joins (intermodal transfer links) were added to the network database to connect the individual modal networks together for routing purposes. An intermodal terminals database and a number of terminal transfer models were developed at RITA/BTS to identify likely transfer points for freight. An algorithm was used to find the minimum impedance path between a shipment's origin ZIP Code to the transfer point and then from the transfer point to the destination ZIP Code. The cumulative length of the spatial joins plus links on this path provided the estimated distances used in CFS mileage computations.

The mileage for an export shipment was calculated between the centroid of U.S. origin ZIP Code and the border crossing on the path of minimum impedance to the foreign destination country (foreign city in the case of Canada and Mexico). For all exports, a POE was found, be it seaport, airport, or border crossing, if not already respondent-provided. However, only the portion of mileage measured within U.S. borders was included as domestic mileage in the CFS estimates.

Methodological Changes from Past Commodity Flow Surveys

Improvements in routing logic - particularly for highway, railway, and airway - were built into the GeoMiler software. Through the use of GeoMiler, distance calculations for freight transportation have been refined to better estimate the actual shipment mileage. In particular, GeoMiler introduced an overall concept change in algorithm for:

- Highway Routing

To estimate highway mileage, GeoMiler considered the functional class of highway so that the "best path" was the quickest path, based on the likely use of interstate and other major roadways, and not necessarily the shortest path. The "quickest path" algorithms in terms of travel time incorporated the following hierarchical functional class of highway:

1. Interstate route;
2. U.S. route;
3. State route;
4. County or other local route.

Hence, the 2007 highway model favored the selection of the higher-order routes (Interstate) rather than lower-order routes (State and County), which provided a more realistic path for freight movement via highway.

The use of these selection criteria, coupled with a more extensive highway network, produced slightly higher (an average of about 3 percent) mileages on highway shipments of distances less than 300 miles.

- Railway Routing

To estimate railway mileage, GeoMiler selected a "single best path" from those calibrated with route density information obtained from sampled 2005 rail waybills, assigned a specific railroad company at shipment origin, and considered ownership, trackage rights, and interlining (the transfer from one railroad company's trackage network to that of another).

The use of these selection criteria produced slightly higher (an average of about 3 percent) mileages on railway shipments.

- Waterway Routing on Export Shipments

The mileage estimate on an export shipment via airway or waterway included the travel distance over domestic airspace, or on domestic waters, up to the U.S. territorial border. To obtain domestic mileage for export shipments measured in the 2007 CFS, mileages were calculated to the U.S. border for ALL modes of transportation on ALL export shipments.

For waterway exports via inland waterways (for example, the Mississippi River), the mileage calculation was continued from an inland water POE (such as St. Louis) to a coastal POE (such as New Orleans), and this extra inland waterway mileage was included in the total domestic mileage for this shipment.

The use of these selection criteria on waterway exports via inland waterways resulted in negligible changes to mileages on inland waterways.

For waterway exports via the Great Lakes (Lakes Erie, Huron, Michigan, Ontario, Superior), the mileage calculation was continued from a Great Lakes POE (such as Chicago, Cleveland, Duluth) to the line of demarcation between the United States and Canada (drawn within each of the Great Lakes except Michigan), and this extra Great Lakes mileage was included in the total domestic mileage for this shipment.

The use of these selection criteria on waterway exports via the Great Lakes produced much higher (an average of about 15 percent) mileages on Great Lakes waterways.

- Airway Routing on Both Domestic and Export Shipments

To estimate domestic airway mileage, GeoMiler employed a "single best path" algorithm from the three airports closest to the origin ZIP Code to the three airports closest to the destination ZIP Code, calibrated with 2005 air route information provided by RITA/BTS/OAI. As in the past, to be acceptable, an airway routing must generate at least twice as many airway miles as highway miles (at least 2:1 ratio of air:truck miles) in order to reach the destination.

Consequently, this selection benchmark in the 2007 CFS chose the most likely air route from those routes that were non-stop (direct) from airport facilities with higher cargo lifts (weight transported between two airports), based on the OAI air cargo data.

For airway exports via inland airports (such as Denver, Memphis), the mileage calculation was continued from an inland air POE to a coastal point on the U.S. landmass (where the air flight path to a foreign country intersected with the U.S. territorial border), and this extra airway mileage was included in the total domestic mileage for this shipment.

The use of these selection criteria on both domestic airway and airway exports via inland airports, coupled with a more extensive airway network, produced much higher (an average of about 12 percent) mileages on airways.

- Routing in Alaska

Much of the State of Alaska was inaccessible by any mode of transportation, except "bush" airplanes. A bush airplane is a small aircraft that usually carries no more than four people, including the bush pilot. For the 2007 CFS, a network of mini-airports, more extensive than that used previously, was incorporated into intrastate travel within Alaska to accommodate "short-hop" flights where no established roads existed, especially in cases where the respondent reported a mode of highway.

- Pipeline mileage

For most of the pipeline shipments, the respondents reported the shipment destination as a pipeline facility on the main pipeline network. Therefore, for the majority of these shipments, the resulting mileage represented only the access distance through feeder pipelines to the main pipeline network, and not the actual distance through the main pipeline network. Due to restrictions on use of the national pipeline network, GeoMiler calculates great circle distance (GCD) for all pipeline shipments.

Back to top

Industry Coverage

The 2007 Commodity Flow Survey (CFS) covers business establishments with paid employees that are located in the United States and are classified using the 2002 North American Industry Classification System (NAICS) in mining, manufacturing, wholesale, and selected retail and services trade industries, namely, electronic shopping and mail-order houses, fuel dealers, and publishers. Additionally, the survey covers auxiliary establishments (i.e., warehouses and managing offices) of multi-establishments companies. For the 2007 CFS, an advance survey (pre-canvass) of approximately 40,000 auxiliary establishments was conducted to identify auxiliary establishments with shipping activity. Surveyed establishments that indicated undertaking shipping activities and the non-respondents to the pre-canvass were included in the CFS sample universe.

Establishments classified in transportation, construction, and most retail and services industries are excluded from the survey. Farms, fisheries, foreign establishments, and most government-owned establishments are also excluded.

In-scope industries for the 2007 CFS were selected based on the 2002 version of the NAICS, while the industries included in the 2002 CFS were selected based on the 1997 version of the NAICS. However, the industries in the 1993 CFS and the 1997 CFS were selected based on the 1987 Standard Industrial Classification System (SIC) and, although attempts were made to maintain similar coverage among the SIC based surveys (1993 and 1997) and the NAICS based surveys (2002 and 2007), there have been some changes in industry coverage due to the conversion from SIC to NAICS. Most notably, coverage of the logging industry changed from an in-scope Manufacturing (SIC 2411) to the out-of-scope sector of Agriculture, Forestry, Fishing, and Hunting under NAICS 1133. Also, publishers were reclassified from Manufacturing (SIC 2711, 2721, 2731, 2741, and part of 2771) to Information (NAICS 5111 and 51223) and were excluded in the 2002 CFS. The 2007 CFS, however, includes publishers and retail fuel dealers.

The NAICS industries covered in the 2007 CFS are listed in the following table:

NAICS code	Description
212	Mining (Except Oil and Gas)
311	Food Manufacturing
312	Beverage and Tobacco Product Manufacturing
313	Textile Mills
314	Textile Product Mills
315	Apparel Manufacturing
316	Leather and Allied Product Manufacturing
321	Wood Product Manufacturing
322	Paper Manufacturing
323[1]	Printing and Related Support Activities (except 323122)
324	Petroleum and Coal Products Manufacturing
325	Chemical Manufacturing
326	Plastics and Rubber Products Manufacturing
327	Nonmetallic Mineral Product Manufacturing
331	Primary Metal Manufacturing
332	Fabricated Metal Product Manufacturing
333	Machinery Manufacturing
334	Computer and Electronic Product Manufacturing
335	Electrical Equipment, Appliance, and Component Manufacturing
336	Transportation Equipment Manufacturing
337	Furniture and Related Product Manufacturing
339	Miscellaneous Manufacturing
423	Wholesale Trade, Durable Goods
424	Wholesale Trade, Nondurable Goods
4541	Electronic Shopping and Mail-Order Houses
45431	Fuel Dealers
4931[2]	Warehousing and Storage
5111	Newspaper, Periodical, Book, and Directory Publishers
51223[3]	Music Publishers
551114[4]	Corporate, Subsidiary, and Regional Managing Offices

[1] Excludes Pre-Press Services (NAICS 323122)

[2] Includes only captive warehouses that provide storage and shipping support to a single company. Warehouses offering their services to the general public and other businesses are excluded.

[3] For tabulation and publication purposes, NAICS 51223 is grouped with NAICS 5111.

[4] Includes only those establishments in NAICS 551114 with shipping activity.

Other industry areas that are not covered, but may have significant shipping activity, include agriculture and government. For agriculture, specifically, this means that the CFS does not cover shipments of agricultural products from the farm site to the processing centers or terminal elevators (most likely short-distance local movements), but does cover the shipments of these products from the initial processing centers or terminal elevators onward.

Back to top

Shipment Coverage

The CFS captures data on shipments originating from select types of business establishments located in the 50 states and the District of Columbia. The data do not cover shipments originating from business establishments located in Puerto Rico and other U.S. possessions and territories. Shipments traversing the U.S. from a foreign location to another foreign location (e.g., from Canada to Mexico) are not included, nor are shipments from a foreign location to a U.S. location. However, imported products are included in the CFS at the point that they leave the importer's initial domestic location for shipment to another location. Shipments that are shipped through a foreign territory with both the origin and destination in the U.S. are included in the CFS data. The mileages calculated for these shipments exclude the international segments (e.g., shipments from New York to Michigan through Canada do not include any mileages for Canada). Export shipments are included, with the domestic destination defined as the U.S. port, airport, or border crossing of exit from U.S. See the "Mileage Calculation" section for additional detail on how mileage estimates were developed.

Back to top

Overview

The sample for the 2007 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 groups of four 1-week periods (reporting weeks) within the survey year, and the third-stage sampling units were shipments.

Back to top

First Stage - establishment selection

Sampling frame

To create the first-stage sampling frame, a subset of establishment records (as of August 2006) was extracted from the Census Bureau 's Business Register. The Business Register is a database of all known establishments located in the United States or its territories, and an establishment is a single physical location where business transactions take place or services are performed. Establishments located in the United States, having nonzero payroll in 2005, and classified in mining (except oil and gas extraction), manufacturing, wholesale, electronic shopping and mail order, fuel dealers, and publishing industries, as defined by the 2002 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. Establishments classified in forestry, fishing, utilities, construction, transportation, and all other retail and services industries were not included on the sampling frame. Farms and government-owned entities (except government-owned liquor stores) were also excluded from the sampling frame. The resulting frame comprised approximately 754,000 establishments as listed in the table below.

Trade Area	**Establishments**
Mining	6,789
Manufacturing	327,826
Wholesale	356,477
Retail	25,190
Services	22,539
Auxiliaries	14,878
Total	753,699

For each establishment, sales, payroll, number of employees, a six-digit NAICS code, name and address, and a primary identifier were extracted, and a measure of size was computed. The measure of size was designed to approximate an establishment's annual total value of shipments for the year 2004.

All of the establishments included on the sampling frame had state, county, and place geographic codes, which were used to assign each establishment to one of the 73 metropolitan areas (MAs) defined as a combination of the metropolitan statistical areas (MSAs), combined statistical areas (CSAs) and states. Establishments not located in an MA were assigned to the balance of the state.

Stratification

The sampling frame was stratified by geography and industry. A particular geographic-by-industry combination defined a *primary stratum*. Geographic strata were defined by a combination of the 50 states, the District of Columbia, and 65 metropolitan areas (MAs) based on their population and importance as transportation gateways. All other MAs were collapsed with the non-MAs within the state into Rest of State (ROS) strata. When an MA crossed state boundaries, size of each part of the MA was considered relative to the MAs total measure of size when determining whether or not to create strata in each state in which the MA was defined. Six MAs had strata in two or more states.

The industry strata were determined as follows. Within each of the geographic strata, 48 industry groups were defined based on the 2002 NAICS:

- three mining (four-digit NAICS);
- 21 manufacturing (three-digit NAICS);
- 18 wholesale (four-digit NAICS);
- two retail (NAICS 4541 and 45431);
- one services (NAICS 5111 and 51223 combined) and
- three auxiliary (combinations of NAICS 4931 and 551114).

If a three or four digit NAICS industry contributed at least 4% of the total value (based on sampling measure of size) or tonnage (based on 2002 CFS data) for the geographic stratum or the nation, it was designated as a *do not collapse* industry stratum within the geographic stratum. Industries not meeting this level of activity within a geographic stratum were grouped with other similar industries. The remaining industry strata were collapsed to form at most 10 *collapsed* industry strata within each geographic stratum.

The method used to collapse the remaining strata, used 2002 CFS data as input to a Classification and Regression Tree (CART) procedure that related industries with commodities. The terminal nodes from the CART procedure were then grouped using a hierarchical clustering algorithm. Using the results from the hierarchical clustering algorithm, some of the clusters were manually regrouped to arrive at the final industry clusters.

To produce better estimates of the shipment of hazardous materials for 2007, a total of 160 strata targeting HAZMAT shippers were created. Using 2002 CFS data, the 6-digit NAICS industries that accounted for a large proportion of the estimated total value and/or total tonnage for six groups of hazardous materials was identified. These included ammonium nitrate, ethanol, explosives, hydrogen, toxic by inhalation, and all other miscellaneous hazardous materials.

The treatment of auxiliary establishments was modified for 2007 to take advantage of the data collected through the advance survey. For auxiliaries that responded to the advance survey and were considered to be shippers, 123 strata were created, one in each geographic stratum, combining both NAICS 4931 and 551114. Two national strata for auxiliary establishments were also created for those that did not respond to the advance survey - one stratum for non-responding warehouses (those classified in NAICS 4931) and one stratum for non-responding management offices (NAICS 551114).

The table below summarizes the primary stratification of the CFS sampling frame. Of the 2,745 primary strata, 232 were designated as take-all strata because of the small number of establishments in the stratum and/or their importance.

Primary Strata	**Number**
Do Not Collapse	1,306
Collapsed	1,154
Auxiliaries (Advance Survey responders)	123
Auxiliaries (Advance Survey non-responders)	2
HAZMAT	160
Total	2,745

Sample size and allocation

Sample sizes were computed to meet coefficient of variation (CV) constraints on estimated value of shipments totals for each primary stratum. A CV of 1.5% on the estimated total value of shipments was used for each primary stratum because it produced total sample sizes of approximately 100,000 establishments.

The primary constraints were budget related, which are translated into an approximate fixed sample size for the survey. The goal of the design was to allocate this fixed total sample size in a statistically efficient manner. The CV constraints were primarily used as a tool to allocate more of the sample to more important strata. It was assumed that the cost of data collection would not vary by stratum. Maximum sampling weight and minimum sample size constraints were also imposed. For the CFS designs, the maximum first stage sample weight was set to 100 and the minimum sample size to 2 establishments per stratum.

The procedure for determining sampling parameters was an iterative computerized process. The sample design programs used in the process are part of a group of generalized programs that have been modified to accommodate the needs of the survey, but use common methods such as the Dalenius and Hodges cumulative sqrt(f) procedure, Neyman allocation, and similar rules for determining acceptable designs.

For each (non-take all) primary sampling stratum, the survey designer specified as input to a Generalized Univariate Stratification (GUS) program:

- desired number of bins (for a frequency distribution used in the Dalenius and Hodges' cumulative sqrt(f) procedure)
- desired number of size strata
- desired number of certainty companies
- desired coefficient of variation for total value of shipments

- maximum sampling weight
- minimum sample size

Once designs were determined for each of the primary strata, the information from these designs was used as input to a program that attempted to more efficiently allocate the sample to meet the desired CV on each primary stratum and also determine the sample sizes needed to meet a national level constraint. Designs with a national level constraint tend to allocate more samples to the larger states so there is a trade off between better national estimates and the quality of the more detailed geographic estimates. For the 2007 CFS, a design with a primary strata CV of 1.7% and a national CV of 0.036% was chosen. The final first stage sample size was 102,369 establishments.

Back to top

Second Stage - reporting week selection

The frame for the second stage of sampling consisted of 52-weeks from January 6, 2007 to January 4, 2008. Each establishment selected into the 2007 CFS sample was systematically assigned to report for four reporting weeks-one in each quarter of the reference year. Each of the 4-weeks was in the same relative position of the quarter. For example, an establishment might have been requested to report data for the 5th, 18th, 31st, and 44th weeks of the reference year. In this instance, each reporting week corresponds to the 5th week of each quarter. Prior to assignment of weeks to establishments, the selected sample was sorted by primary stratum (state x metropolitan area x industry) and measure-of-size.

Back to top

Third Stage - shipment selection

For each of the four reporting weeks in which an establishment was asked to report, the respondent was requested to construct a sampling frame consisting of all shipments made by the establishment in the reporting week. Each respondent was asked to count or estimate 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, the respondent was asked to select a systematic sample of the establishment's shipments and to provide information only about the selected shipments. If an establishment made *40 or fewer* shipments during that week, the respondent was asked to provide information about *all* of the establishment's shipments made during that week (i.e., no sampling was required).

Back to top

Last Revised: January 25, 2017