MEASURING RETAIL E-COMMERCE SALES

LaTasha I. Austin, Carol S. King, Christopher Pece, and Judith O'Neil, Bureau of the Census
LaTasha Austin, Bureau of the Census, SSSD, Washington, DC 20233

Key Words: E-commerce, imputation, estimation, data quality

Introduction

The growth of electronic commerce (e-commerce) in the past several years has been phenomenal. The value of internet purchases of goods and services by individuals and businesses has been measured by many private forecasters. On March 2, 2000 the Census Bureau released the U.S. Federal Government's first official measure of retail e-commerce sales for the Fourth Quarter 1999 Holiday season (October, November, December 1999). E-commerce sales are sales of goods and services over the Internet, an extranet, electronic data interchange (EDI), or other online system, where payment may or may not be made online. This paper[1] provides an overview of the coverage, collection, imputation, and estimation methods used for the e-commerce estimates, presents results of the survey, and discusses issues related to the quality of the results and plans for releasing future e-commerce estimates.

1. Sample Selection

The Retail e-commerce sales are estimated from the same sample used in the Monthly Retail Trade Survey (MRTS) to estimate U.S. retail sales. The sampling frame for the MRTS was extracted from the Standard Statistical Establishment List (SSEL). The SSEL is maintained and updated regularly by the Census Bureau. It includes administrative receipts and information from the Economic Census, the Internal Revenue Service (IRS), and several other resources. The SSEL also contains the payroll Employer Identification Number (EIN) and company affiliation for each employer business establishment location in the United States. The IRS assigns the payroll EINs which are used by employer businesses to report Social Security (SS) payments for its employers. Companies with more than one establishment may have one or more EINs.

Using information from the SSEL, retail company sampling units are formed. A stratified simple random sampling method is used to select over 13,500 firms whose sales are then weighted and benchmarked to represent the complete universe of over two million retail firms. These sampling units are stratified by major kind of business (MKB) and estimated sales. Each company with sales above a preset sales cutoff for its MKB is selected into the sample with certainty. Approximately 3,000 of the selected survey units are certainty. Retail companies not selected as certainty are disaggregated into their EINs based on information from the SSEL. The EINs are then stratified by MKB and sales and a simple random sample of EINs is selected from each stratum.

The MRTS sample is updated on a quarterly basis to account for new EINs identified as active by the IRS. To select a sample of birth EINs for addition to the MRTS mailout, a two-phase procedure is used. In the first phase, births are stratified by kind of business (KB) and size (expected payroll or employment). A relatively large stratified simple random sample is selected and canvassed to obtain sales measures of size and more detailed KB codes. Using this information, the EINs are subjected to a second phase of sampling with overall probabilities equivalent to those used in the initial MRTS sample. Since companies and EINs engaged in e-commerce retailing are a subset of all retail companies and EINs, the sample selection and maintenance procedures noted above apply to survey units engaging in e-commerce.

In addition to the procedures noted above, to ensure adequate coverage of retail e-commerce sales, we identified businesses known to be engaged in e-commerce and determined their sales volume. Using this information we assessed whether these businesses were represented appropriately in our sample. In a few instances, we adjusted the sample to be more representative of retailers engaged in e-commerce.

The retail firms include businesses such as building material dealers, new car dealers, furniture stores, mail order, eating and drinking places, grocery stores, apparel

[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.

stores, and companies selling merchandise for personal and household consumption. The MRTS excludes firms such as financial brokers and dealers, and ticket sales agencies, and therefore these types of firms are not included in retail e-commerce sales estimates.

The characteristics of the seller determines not only how we tabulated our data, but it also determines where we tabulated our data. For example, firms primarily selling via the internet are canvassed in the Standard Industrial Classification (SIC) 5961, Catalog, Electronic Shopping and Mail Order Houses. However, if e-commerce sales are not the primary source of receipts, the data is most likely tabulated in the industry in which its traditional retail stores are classified such as toys, books, or general merchandise.

As a first step in collecting e-commerce sales, a letter was sent to each reporting unit in August 1999, describing the e-commerce study and the information needed before the updated monthly form was mailed for October 1999 requesting the e-commerce sales. Reporting units were classified as one of the following depending on their response to the screener:

- Units that currently have e-commerce sales;
- Units that currently do not have e-commerce sales but plan to during the 4th quarter;
- Units that currently do not have e-commerce sales and do not plan to have any during the 4th quarter.

About 14 percent of the reporting units mailed responded that they either had e-commerce sales or planned to during the 4th quarter. The screener identified MRTS survey units to target for e-commerce sales data collection. These units were mailed a form to which e-commerce questions were added. Thus both total monthly sales and e-commerce sales were collected for each of the months October, November, and December 1999.

2. Editing Methodology

Those reporting units that currently have e-commerce sales were to report their e-commerce sales as a percent of total sales or as a dollar volume. Some basic checks and computations were performed to edit the data.

- If a unit reported the percent but not e-commerce sales then,

  *e-commerce sales = percent * monthly sales*

- If a unit reported e-commerce sales but not percent and e-commerce sales < monthly sales,

  percent = e-commerce sales / monthly sales

- If the e-commerce sales was greater than the monthly sales, the e-commerce sales was set for imputation.

- If the unit reported both e-commerce sales and percent then

  *computed e-commerce = percent * monthly sales*

- If |computed e-commerce sales – reported e-commerce sales|>50 then both the percent and the e-commerce sales were set for imputation.

Having no basis as to what to consider an outlier, we decided to not identify specific outlier e-commerce data. Instead, specific ratios were computed and the units were ranked based on the size of the ratio. Analysts could then review as many of the units as time permitted.

We computed the following ratios for each reporting unit, creating a list for each type of ratio computed:

- Current Month (CM) e-commerce sales / CM monthly sales
- CM e-commerce sales / Previous Month (PM) e-commerce sales
- (CM e-commerce sales / CM monthly sales) - (PM e-commerce sales / PM monthly sales)
- |(CM e-commerce sales / CM monthly sales) - (PM e-commerce sales / PM monthly sales)|

The units were sorted by SIC and descending ratio for each listing.

3. Imputation Methodology

The e-commerce sales and percent of total sales were imputed on a case by case basis, for nonresponse cases, for cases who reported e-commerce sales > total sales, and for cases who had more than a difference of $50 (*in E-commerce sales*) between their reported e-commerce sales and our computed e-commerce sales. The following sets of ratios were computed:

Ratio 1: $\sum(W_i * \text{e-commerce sales}) / \sum(W_i * \text{monthly sales})$, where W_i equals the monthly sampling weight. This ratio included all units that reported e-commerce sales and was computed separately for units with weight = 1.000

(certainty units) and for units with weight > 1.000 (noncertainty units).

Ratio 1a: The same ratio as above but also including units that answered on the screener that they had no e-commerce sales and did not plan to have any during the quarter.

Ratio 2: (Current month (CM) sales) / (Previous month (PM) sales). This was computed from each firm's CM and PM sales data.

Ratio 3: $\sum(W_i * \text{CM e-commerce sales}) / \sum(W_i * \text{PM e-commerce sales})$. This ratio was also computed separately for units with weight = 1.000 (certainty units) and for units with weight > 1.000 (noncertainty units) and included data from units for which both CM and PM e-commerce sales were reported.

All ratios except for Ratio 2 were computed at the 4-digit Standard Industrial Classification (SIC) levels.

The resulting group of ratios in each category are referred to as the **imputation base.**

Imputation took into account the response to the screener, kind of business, and sales size. The methodology for each month was as follows:

3.1 October 1999

Condition	Imputation of E-commerce Sales
Reporting units stating in the screener that they had no e-commerce sales	0
Reporting units stating they had e-commerce sales but did not report any	CM monthly sales * R1[2]
Reporting units that did not respond to screener or were not mailed a screener	CM monthly sales * R1a
Reporting units only having e-commerce sales	CM monthly sales

[2] R1, R1a, R2, and R3 correspond to Ratio 1, Ratio 1a Ratio 2, and Ratio 3, respectively.

3.2 November and December 1999

Category	Condition	Imputation of E-commerce Sales
1	Reporting units stating in the screener that they had no e-commerce sales	0
2	Reporting units stating they had e-commerce sales but did not report any	CM monthly sales * RI
		PM e-commerce sales * R2
		PM e-commerce sales * R3
3	Reporting units that did not respond to screener or were not mailed a screener and were not births	CM monthly sales * R1a
		PM e-commerce sales * R2
		PM e-commerce sales * R3
4	Reporting units only having e-commerce sales	CM monthly sales
5	Births	CM monthly sales * R1

Tables of e-commerce sales were created based on the types of imputation methodology used at the four-digit SIC level. These tables included counts and percentages for the number of units tabulated by reported versus imputed along with the corresponding dollar volume. The tabulations were used to determine which imputation methodology provided the "best" estimate of e-commerce sales for categories 2 and 3. In general, the imputed estimate using Ratio 1 was the largest estimate, the estimate using Ratio 2 was the smallest estimate, and the estimate using Ratio 3 was in between. We chose Ratio 3, because we felt that it best reflected the month-to-month change that occurred in the e-commerce sector and also took into consideration the belief that units of like size behaved in a similar manner. This method yielded

an imputation rate of approximately 20 percent for the total e-commerce sales estimate. The corresponding imputation rate for total sales was 26 percent. We suspect the imputation rate was smaller for the e-commerce estimate because those firms that tended to have e-commerce sales are the larger firms. The larger firms normally are the best respondents.

4. Estimation

For each month of the quarter, estimates were obtained by aggregating weighted e-commerce sales to which the carry forward factors reflecting the relationship of the 1997 Retail Census and the 1997 Annual Retail Trade Survey (ARTS) estimates were applied. These carry forward factors account for nonemployers and differences in the employer portion such as classification or other coverage problems. To obtain the quarterly estimates, the monthly benchmarked estimates were summed. The percent of e-commerce sales to total retail sales was also computed. Margins of error were computed for both the level estimate and the percentage. Using the margins of error, confidence intervals were also obtained.

5. Private Sector Estimates of E-commerce Sales

How private sector forecasters arrived at their estimates varied quite a bit. We researched 13 of these forecasters, reviewing their estimate of holiday sales forecast, forecast period, type of estimate, survey unit, sample size, coverage, and survey methodology. Level estimates ranged from $3.2 billion to $13 billion and percent estimates from 1 to 3 percent of total retail sales. Except for three of them, the time period covered by these estimates was 4th quarter 1999. Many of these estimates were obtained from consumer-based samples, with the number of consumers varying from 300 to 770,000. The number of firms contacted for the business based estimates ranged from 30 to 300 firms. These estimates covered a variety of products and services such as apparel, computer goods, food and wine, gift, and travel and financial services.

6. Data Quality Issues

Quality of the e-commerce data are affected by such things as classification, timing of the introduction of new businesses, nonresponse, and the use of a survey designed to report total sales to collect e-commerce sales. A survey unit reporting e-commerce sales can be tabulated in either SIC 5961, Catalog, Electronic Shopping and Mail Order Houses or in the SIC in which its traditional stores are classified. Where the unit is classified not only effects the estimate at the particular level but also affects imputation. Imputation cells are created based on SIC. We also face the issue of survey units selected under one SIC but later found to be in another inscope SIC. We have looked into a method where we measure the effect of the particular unit on the total sales and e-commerce sales of its old (incorrect) SIC and its new (correct) SIC.

The sample for MRTS was drawn to provide a statistically sound estimate for total sales that met certain CV requirements at various KB levels. It was not designed to provide the same accuracy for e-commerce sales. To overcome this we found it necessary to review the representation of businesses primarily doing e-commerce transactions in our sample. As mentioned before there were some survey units added as a result of our review. It is important that we add to the sample these types of units as quickly as possible. We attempt to identify them by searching websites that provide number of hits by site, reviewing articles in newspapers and magazines concerning these types of businesses, and using our own SSEL to identify .com, .net and .org businesses. This SSEL search is done on a monthly basis allowing analysts to identify businesses sooner than our quarterly processing of births.

Closely related to the above is the issue of representation of total sales versus e-commerce sales. A survey unit with a weight of 100 may well represent 99 other units with respect to total sales, but may not in terms of e-commerce sales. This may be especially evident where the unit has a majority or all of its sales generated through e-commerce.

The MRTS is a voluntary survey. Because of this, a number of survey units refuse to answer the questionnaire. Section 3 describes how we impute for these units as well as for other units not mailed a survey form (verified refusals, units that are out-of-business but tabulated to represent births, and units in out-of-scope KBs). Some of the challenges we face for imputation include the identification of survey units with unusual CM-to-PM e-commerce ratios. Currently this review makes use of the listings mentioned in Section 2. We are looking to automate this identification and flag the unit from inclusion in the imputation base after we get data collected for additional quarters. We will use this data to come up with parameters to identify outliers. We have also encountered having very few units in the base for some SICs. We are looking into collapsing some of our imputation cells by SIC or across our certainty/noncertainty boundary.

7. Results and Future Plans

On March 2, 2000, the Census Bureau reported an estimate of $5.3 billion ($± 0.3 billion) in retail e-commerce sales for the preliminary 4th quarter of 1999 (October through December). This estimate accounted for .64 percent of the total retail sales estimate of $821.2 billion (±0.05 percent) for the quarter. The decision was made to continue the e-commerce survey. Recently, the Census Bureau released the preliminary 1st quarter 2000 estimate of $5.3 billion in retail e-commerce sales. During this time, a revised estimate of $5.2 billion in retail e-commerce sales for the 4th quarter was released as well. E-commerce sales accounted for 0.70 percent of total sales in the first quarter 2000 and 0.63 percent of total sales in the fourth quarter 1999.

The Census Bureau is collecting 1998 and 1999 e-commerce sales data in the 1999 annual survey of manufacturing; wholesale trade; retail trade; food services and accommodations; and information, transportation, business, professional, health care, and personal services.

The Census Bureau will continue to monitor how the imputation methodology for the MRTS estimates work for the e-commerce estimates. In addition, research on other methodology will be conducted to ensure sound estimates. Our research will continue in the areas of timely identification of new e-commerce businesses and investigating alternate editing, imputation, and estimation methods.

References:

Burton, J. (1997), "Monthly Retail and Wholesale Trade Survey Imputation," Memorandum for the Record August 13, 1997.

Detlefsen, R. & McElhatten, M. (2000), "Retail E-commerce Sales Estimates- Methods and Results," Memorandum for the Record.

King, C. S. and Austin, L. (2000), "Edits and Imputation for the Preliminary 4th Quarter 1999 E-commerce Sales," Memorandum for the Record March 21, 2000.