

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Ottawa, Canada, 16-18 May 2005)

Topic (i): Editing administrative data and combined data sources

**THE USE OF ADMINISTRATIVE DATA IN THE ANNUAL SURVEYS OF RETAIL,
WHOLESALE, AND SERVICES¹**

Supporting Paper

Submitted by the U.S. Census Bureau, United States of America²

I. INTRODUCTION

1. Administrative records play a significant part in the production of economic statistics by the U.S. Census Bureau. These records aid in the construction of frames for censuses and surveys, improvement in measures of size for sampling, identification of new businesses and firms that no longer have business activity, and imputation for non-response to a census or survey mailing.

2. The Census Bureau receives administrative record data from other government agencies including the Internal Revenue Service (IRS), the Social Security Administration (SSA), and the Bureau of Labor Statistics (BLS). These data are the foundation for the construction and maintenance of the Census Bureau's Business Register (BR). We have continued to look at various ways to use the administrative record data to improve statistical programs while reducing burden on those businesses asked to respond to our censuses and surveys. The Census Bureau has conducted a number of studies that have looked specifically at how administrative receipts, inventory, and expenses can be used in our annual retail, wholesale, and services surveys.

3. The Annual Retail Trade Survey (ARTS) survey collects sales, inventory, purchases, accounts receivables, and e-commerce (for selected kinds of business) from firms primarily engaged in retail business activity. The Annual Trade Survey (ATS) survey collects sales, inventory, purchases, e-commerce, and expenses (for non-merchant wholesalers) from firms primarily engaged in wholesale business activity. The Service Annual Survey collects receipts, expenses, e-commerce, and detailed receipts and expense lines for those firms primarily engaged in services business activity. The data are collected in six surveys: SAS-Computer, SAS-Health, SAS-Information, SAS-General, SAS-Transportation, and SAS-Finance.

4. These surveys are composed of units selected from a probability based stratified random sample for employer businesses supplemented with administrative data for non-employer businesses.

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.

² Prepared by Carol S. King (Carol.Steadman.King@census.gov).

5. Before the discussion of the administrative data studies, it will be helpful to understand the components of the annual surveys.

- (a) The *certainty component* consists of employer business units that are self-representing (sample weight = 1) because they are expected to have a large effect on the precision of the estimates. These businesses include companies identified by a Census assigned code called an alpha and Federal Employer Identification Numbers (EINs) associated with multi-establishments firms or single units (SUs). The IRS issues each employer business a unique nine-digit as the primary identifier for single establishment businesses and as a secondary identifier for multi-establishment businesses.
- (b) The *non-certainty component* consists of a sample of EINs associated with companies not included in the certainty component. For a given company with paid employees, the IRS issues one or more EINs for payroll tax withholding purposes. Thus, a given EIN is an aggregation of establishments and represents a particular part of its parent company. We refer to two types of EIN sampling units – *single unit EINs* and *multi unit EINs*. A single unit EIN is associated with a company comprised of *only one* establishment, while a multiunit EIN is associated with a company comprised of *more than one* establishment.
- (c) The *non-employer component* consists of firms with no paid employees. Data from this component are obtained from administrative records (IRS tax returns) and are identified and tabulated only for retail and service industries.

II. THE 1998 ADMINISTRATIVE DATA STUDY

6. Konschnik, et al. (1998) evaluated the use of administrative receipts of small SU EINs in place of survey responses to produce estimates from the 1996 ARTS, ATS, and SAS. This study looked at the impact of replacing receipts collected on report forms with annual tax return data for single establishment firms in these annual surveys. In particular, we looked at replacing receipts data for all non-certainty SUs (weight >1) and then for all non-certainty SUs below specified payroll cut-offs (in a manner consistent with the Census Bureau's Economic Census.)

7. This study showed that administrative sales/receipts data could be used as reasonable substitutes for sales/receipts and revenue data reported by SU EINs with annual payroll below specified payroll cut-offs, while reducing both data collection costs and reporting burden on these small businesses. Only slight changes were found in the overall estimates of retail sales, wholesale sales, and service revenue.

A. Use of Administrative Data in the 1999 SAS

8. Beginning with the 1999 survey year, in lieu of mailing forms to all SU EINs canvassed in SAS, the Census Bureau chose to use available administrative receipts data for small SU EINs, if few data items other than revenue were to be collected. These small SU EINs were classified in particular NAICS industries covered by SAS that included Couriers and Messengers (NAICS 492); Warehousing and Storage (NAICS 493); Rental and Leasing Services (NAICS 532); Professional, Scientific, and Technical Services, excluding Computer Systems Design and Related Services (NAICS 54\5415); Administrative and Support and Waste Management and Remediation Services (NAICS 56); Social Assistance (NAICS 624); Arts, Entertainment, and Recreation (NAICS 71); and Other Services (NAICS 81).

9. We decided to mail 1999 SAS forms to all SU EINs canvassed in SAS and classified in Advertising Agencies (NAICS 541810), Travel Agencies (NAICS 561510), or Tour Operators (NAICS 561520) although only a few items were to be collected for them. For these industries, Konschnik, et al. (1998) reported non-ignorable changes in total revenue estimates when administrative receipts data for small SU EINs were substituted for reported revenue data. We believe that administrative receipts data on these EINs include commissions, which should be excluded from revenue data reported in SAS.

10. To identify SU EINs to be withheld from the 1999 SAS mailings, we first determined NAICS-based annual payroll cut-offs using 1997 Economic Census data. We then withheld from the mailings any SU EIN that had 1999 annual payroll less than the payroll cut-off corresponding to the NAICS industry in which the EIN was classified.

11. If administrative receipts were not available for a given non-mailed SU EIN, a value for revenue was imputed by applying an estimated revenue-to-payroll regression coefficient to the SU EIN's annual payroll value. The estimated revenue-to-payroll regression coefficient was computed using 1997 Economic Census data from establishments that were classified in the same six-digit NAICS industry as the SU EIN. Values for items other than revenue were imputed using data from reporting units that responded to the 1999 SAS.

Table 1. Contribution of Non-mailed Single unit EINs in 1999 SAS

Survey	Total Number of Units	Number of Units Below Payroll Cut-offs (% of Total Number of Units)	Total Receipts (000,000)	Receipts from Non-mails (% of Total Receipts)	Administrative Receipts for Non-mails (% of Non-mail Receipts)
SAS-G	25,592	5,532 (19%)	1,582,181	147,008 (9.3%)	90,943 (62%)
SAS-T	2,504	99 (4%)	244,844	1,729 (.71%)	1,104 (64 %)

B. Use of Administrative Data in the 2000 ARTS

12. Beginning with the 2000 survey year, we excluded from the ARTS mailings small SU EINs that were classified in Accommodation and Food Services (NAICS 72). As for certain NAICS industries canvassed by SAS, few data items other than sales were collected from reporting units that are classified in this industry. Besides sales, data were collected on e-commerce sales and sales taxes.

13. The method used to identify the particular SU EINs to be excluded from the 2000 ARTS mailings was similar to the one used for SAS. We created NAICS-based annual payroll cut-offs using data from the 1997 Economic Census and compared the 1999 annual payroll for each SU EIN to the EIN's corresponding cut-off. However, we decided to not mail any SU EINs classified in Food Service Contractors (NAICS 722310), because companies with more than one establishment dominated 1997 Census annual payroll for this industry.

14. Of the 1,936 ARTS tabulating units for 1999 that were classified in Accommodation (NAICS 721), 676 units were non-mailed SU EINs for 2000. Of the 3,427 ARTS tabulating units for 1999 that were classified in Food Services and Drinking Places (NAICS 722), 1,384 units were non-mailed SU EINs for 2000. These SU EINs contributed about 15% to the total sales estimate for NAICS 72 in 2001, 13% in 2002 and 12% in 2003. The corresponding administrative data percents were 8%, 7%, and 7%.

15. The methodology used to impute sales for a given non-mailed SU EIN in the 2000 ARTS was similar to the one discussed for SAS except for the incorporation of aggregated monthly retail sales data from the Monthly Retail Trade Survey (MRTS) conducted for 2000. No monthly survey is conducted for service industries. For a given non-mailed SU EIN in the 2000 ARTS, if the number of months in which sales were imputed in the 2000 MRTS was at most three, substitution of aggregated monthly sales data will have a higher priority than other imputation methods, including substitution of available administrative receipts.

III. COMPARING ADMINISTRATIVE INVENTORY TO REPORTED INVENTORY FOR THE 1999 ARTS

16. In early 1999, the IRS began to provide the Census Bureau with beginning and end-of-year inventory data as reported on the 1998 business income tax returns.

17. While comparing administrative end-of-year inventory to the reported 1999 ARTS end-of-year inventory, we observed that, for about half of the SU EINs, the administrative data were roughly 100 times that of the reported data. This most likely occurred because cents were keyed as dollars. A similar result held true for administrative beginning-of-year inventory. We devised an edit on administrative end-of-year inventory that attempted to correct this problem.

18. Edit methods were explored that relied on the determination of acceptable ranges of values about the administrative data and an inventory-to-sales ratio using reported 1999 ARTS data. We decided that an inventory-to-sales ratio might be the best way to edit the data, because the administrative end-of-year inventory data had too much variability to base our edit only on the distribution of these data.

19. To analyse the uses of administrative inventory data for the 1999 ARTS, we completed three imputation runs. The first run used our current methodology, which produced what we call the *tabulated* estimates. The second run substituted administrative inventory data, when available, or other imputed data for non-responding SU EINs. The third run substituted administrative inventory data, when available, or other imputed data for small or non-responding SU EINs.

Table 2. Source of Total Inventory for Imputation Runs

	% Reported	% Imputed	% Using Admin Inventory Data
Imputation Run 1	85.3	14.7	0
Imputation Run 2	85.1	10.7	4.2

20. The estimate at the total level increased by 0.15% and the coefficient of variation (CV) at the total level increased by 2.19% after the administrative data were used for imputation. The table below shows the number of six-digit NAICS industries that had a percentage difference between the tabulated estimates and the estimates using administrative inventory to impute. There were a total of 71 six-digit industries.

Table 3. Number of Industries with the Following Percentage Differences of Estimates

[-10,-5)	[-5,-2)	[-2,-1)	{-1,0)	[0,1)	[1,2)	[2,5)	[5,10)
2	6	8	20	14	9	10	2

21. For all of the six-digit NAICS industries, the percent difference between total inventory estimates was less than 15% in absolute value. Overall, the estimated coefficients of variation (CVs) of total inventory estimates from Run 2 were comparable to those from Run 1.

22. Because substitution of administrative end-of-year inventory data appeared to be a good imputation method, we decided to study the effects of using administrative end-of-year inventory data in lieu of survey responses from small SU EINs in the 1999 ARTS. Using the same methodology for identifying small SU EINs described in Section 3, a third imputation run was performed. In this run, reported sales and inventory data for small SU EINs were deleted and replaced with administrative data or other imputed data. Where available, administrative inventory data were used to impute data for small SU EINs and for SU EINs that did not respond to the 1999 ARTS. The estimate of total inventory for Retail (NAICS 44-45) increased only 0.09% from Run 1 to Run 3. Table 4 shows the percent contribution of reported, administrative, and other imputed data to the total inventory estimate for NAICS 44-45 from the three imputation runs.

Table 4. Source of Total Inventory for Imputation Runs

	% Reported	% Imputed	% Using Admin Inventory Data
Imputation Run 1	85.3	14.7	0
Imputation Run 2	85.1	10.7	4.2
Imputation Run 3	76.4	12.8	10.8

23. Table 5 displays the number of six-digit NAICS industries that have a percent change, within a specified range, between the total inventory estimates produced for the 71 NAICS codes.

Table 5. Number of Industries by Percent Percentage Differences of Estimates

[-10,-5)	[-5,-2)	[-2,-1)	[-1,0)	[0,1)	[1,2)	[2,5)	[5,15)
5	10	8	14	11	5	14	4

24. We decided to incorporate the use of administrative inventory into the annual imputation for SUs in both ARTS and ATS. The results can be seen in the following table.

Table 6. Contribution of Administrative Inventory to the ATS and ARTS

Survey	Survey Year	% Inventory Reported	% Inventory Imputed	% Admin Substituted
ATS	2001	85.8	11.6	2.6
	2002	85.4	12.8	1.8
	2003	88.4	9.7	1.9
ARTS	2001	85.5	10.1	4.4
	2002	86.2	8.7	5.1
	2003	88.5	7.6	3.9

IV. THE USE OF ADMINISTRATIVE EXPENSE DATA FOR SINGLEUNIT ESTABLISHMENTS

25. Beginning with tax year 1998, as with inventory, the Census Bureau received data from the IRS for expenses collected on the tax forms. This study looked at using administrative expenses for SU EINs as an imputation method.

26. For this study survey data from the years 2001, 2002, and 2003 were analysed from the components of the Service Annual Survey (SAS) that collected total expense data.

- For the 2001 and 2002 survey years, this included the general component (SAS-G), the health component (SAS-H), and the information component (SAS-I). Only IRS tax form 990 data were used when comparing to the 2001 and 2002 annual expense data because only tax-exempt EINs were asked to report data.
- For the 2003 survey year, the finance component (SAS-F) and the computer component (SAS-C) began collecting total expense data. Data from these surveys were compared to the

data reported on the IRS's tax forms. All tax data were used to compare to the 2003 survey because both taxable and tax-exempt EINs were asked to report expense data in this year.

27. The analysis showed that after editing expense data it was feasible to use administrative expenses as an imputation method for SU EINs in SAS. Substituting reported expenses with administrative expenses resulted in small effects on the total estimates and standard errors. However, because there is only one year left under our current sample design, we decided not to incorporate the use of administrative expenses for SAS until survey year 2005, when a new sample will be introduced.

V. CONTINUING RESEARCH

28. In addition to completing the research described above, we are also continuing to:
- improve the editing methodology for administrative receipts and inventory;
 - study the degree of agreement between administrative data and reported data for individual sampling units;
 - measure the effect of not mailing survey forms to all SU EINs even if detailed receipts or expense data are collected;
 - study the use of administrative expenses for the non-merchant wholesalers;
 - gain an understanding of the reasons why some selected sampling units with nonzero reported sales and payroll have no administrative receipts on the Business Register;
 - investigate the use of administrative data for multiunit EINs.

References

Gillott, Lora, "The Use of Administrative Expense Data for Singleunit EINs," Draft Memorandum for The Record, January 25, 2005.

Konschnik, Carl A., et al, "The Use of Administrative Records in Current Business Surveys and Censuses," Proceedings of the 1998 Joint Statistical Meetings.

Myers, Anthony L., et al, "The Effects of Using Administrative Data in Place of Annual Business Survey Responses for Small Sampling Units," Proceedings of the Federal Conference on Statistical Methodology, 2001.
