USING ADMINISTRATIVE DATA IN LIEU OF SURVEY RESPONSES FOR SMALL BUSINESSES[1]

Anthony L. Myers, David L. Kinyon, and Carol S. King
U.S. Census Bureau

Abstract

The United States Census Bureau uses data from administrative sources, including the Internal Revenue Service, the Social Security Administration, and the Bureau of Labor Statistics, for important survey processes related to the production of estimates from the Annual Retail Trade Survey, the Annual Trade Survey, and the Service Annual Survey. For these annual business surveys, administrative data are used to create and to update sampling frames, to calculate measures of size used in sample selection, and to impute for survey nonresponse.

This paper documents the results of ongoing research to analyze the potential effects of administrative data on estimates of annual sales and of data items, including beginning-of-year inventory, end-of-year inventory, and annual expenses, for which administrative data have not been previously used for imputation.

Keywords: Business surveys, administrative data, imputation

1. Introduction

The Census Bureau conducts annual surveys of the retail, wholesale, and service sectors, as defined by the 1997 North American Industry Classification System (NAICS). These surveys measure totals and trends that are important to the United States economy. Through the Annual Retail Trade Survey (ARTS) and the Annual Trade Survey (ATS), we collect data such as sales, end-of-year inventory, and value of purchases for retail and wholesale industries. The Service Annual Survey (SAS) collects revenue data, as well as data specific to particular service industries. Estimates from the ARTS and ATS serve as benchmarks for the Monthly Retail Trade Survey and the Monthly Wholesale Trade Survey, respectively.

The samples for these annual surveys consist of three components:

1. The *certainty component* consists of self-representing companies that were given a sampling weight of one because these sampling units were expected to have a large effect on the precision of the estimates. These companies are comprised of one or more *establishments*, where an establishment is the smallest business unit at which transactions take place and payroll and employment records are kept.

---

[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. We thank Ruth E. Detlefsen, Paul S. Hanczaryk, and Richard A. Moore Jr. for their helpful comments.

2. The *noncertainty component* consists of a sample of Employer Identification Numbers (EINs) associated with companies not included in the certainty component. For a given company with paid employees, the Internal Revenue Service (IRS) issues one or more EINs for tax filing and reporting 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 - *singleunit EINs* and *multiunit EINs*. A singleunit 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.

3. The *nonemployer 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.

For more information on the sample design used for ARTS, ATS, and SAS, see Kinyon, et al. (2000).

Data for selected companies and multiunit EIN sampling units may be collected from one or more parts of the unit. We refer to each of these parts as a *reporting unit*. In some instances, for ARTS and ATS tabulation purposes, we split the record for a reporting unit by the reporting unit's major industry groups. We refer to the records used for tabulation as *tabulating units*. For singleunit EIN sampling units, the EINs are both reporting units and tabulating units.

Konschnik, et al. (1998) evaluated the use of administrative receipts of small singleunit EINs in place of survey responses to produce estimates from the 1996 ARTS, ATS, and SAS. This study showed that administrative receipts data could be used as reasonable substitutes for sales and revenue data reported by singleunit EINs with annual payroll below specified payroll cutoffs, 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.

Based on the findings of Konschnik, et al. (1998), we decided to withhold mailing annual report forms to small singleunit EINs classified in retail and service industries for which few data items other than sales or revenue are collected. Sections 2 and 3 discuss the contribution of these singleunit EINs in the 1999 SAS and the 2000 ARTS. Sections 4, 5, and 6 describe research on administrative end-of-year inventory data to investigate the possibility of not mailing 2001 ARTS forms to small singleunit EINs classified in additional retail industries. Section 7 outlines areas for continuing research.

2. Use of Administrative Receipts Data for Small Singleunit EINs in the 1999 Service Annual Survey (SAS)

Beginning with the 1999 survey year, in lieu of mailing forms to all singleunit EINs canvassed in SAS, we chose to use available administrative receipts data for small singleunit EINs, if few data items other than revenue were to be collected. These small singleunit 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). For reporting units classified in these NAICS industries, data are collected on revenue, e-commerce revenue, and expenses.

However, we decided to mail 1999 SAS forms to all singleunit EINs that were canvassed in SAS and were classified in Advertising Agencies (NAICS 541810), Travel Agencies (NAICS 561510), or Tour Operators (NAICS 561520). For these industries, Konschnik, et al. (1998) reported nonignorable changes in total revenue estimates when administrative receipts data on small singleunit EINs were substituted for reported revenue data. We believe that administrative receipts data on these EINs include commissions, which are excluded from revenue data reported in SAS.

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

If administrative receipts were not available for a given nonmailed singleunit EIN, a value for revenue was imputed by applying an estimated revenue-to-payroll regression coefficient to the singleunit 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 singleunit EIN. Values for items other than revenue were imputed using data from reporting units that responded to the 1999 SAS.

Table 1. Contribution of Nonmailed Singleunit EINs in 1999 SAS

NAICS Industry	Total Number of Units	Number of Units Below Payroll Cutoffs (% of Total Number of Units)	% of Total Revenue from Nonmails	% of Nonmail Revenue from Administrative Receipts
492	194	33 (17%)	2%	68%
493	480	66 (14%)	7%	60%
532	1,725	306 (18%)	4%	64%
54\5415	7,451	1,280 (17%)	10%	62%
56	7,003	1,159 (17%)	7%	64%
624	3,415	486 (14%)	7%	51%
71	4,422	1,081 (24%)	12%	61%
81	8,991	1,706 (19%)	12%	60%

For each of the NAICS industries in which singleunit EINs were withheld from the 1999 SAS mailings, the contribution of nonmailed singleunit EINs to the number of reporting units and to the total revenue estimate is given in Table 1. For these industries combined, nonmailed singleunit EINs contributed about 18% to the number of reporting units and about 9% to the total revenue estimate, with administrative receipts data contributing about 61% to total revenue from these EINs.

3. Use of Administrative Receipts Data for Small Singleunit EINs in the 2000 Annual Retail Trade Survey (ARTS)

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

The method used to identify the particular singleunit EINs to be excluded from the 2000 ARTS mailings was similar to the one described in Section 2. We created NAICS-based annual payroll cutoffs using data from the 1997 Economic Census and compared the 1999 annual payroll for each singleunit EIN to the EIN's corresponding cutoff. However, we decided to not mail any singleunit EINs classified in Food Service Contractors (NAICS 722310), because companies with more than one establishment dominated 1997 Census annual payroll for this industry.

Of the 1,936 ARTS tabulating units for 1999 that were classified in Accommodation (NAICS 721), 676 units are nonmailed singleunit 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 are nonmailed singleunit EINs for 2000. We have not determined the contribution of the NAICS 72 nonmailed singleunit EINs to the 2000 sales estimates, because data collection for the 2000 ARTS has not yet been completed. However, in the 1999 ARTS, these singleunit EINs contributed about 15% to the total sales estimate for NAICS 721 and about 18% to the total sales estimate for NAICS 722.

The methodology used to impute sales for a given nonmailed singleunit EIN in the 2000 ARTS will be similar to the one discussed in Section 2, 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 nonmailed singleunit 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.

4. Comparing Administrative Inventory Data to Reported Data for the 1999 Annual Retail Trade Survey

In early 1999, IRS began to provide the Census Bureau with beginning-of-year and end-of-year inventory data as reported on business income tax returns. Because both administrative receipts and end-of-year inventory data could potentially be used to impute data for additional nonmailed

singleunit EINs in the 2001 ARTS, we explored this imputation method using data from the 1999 ARTS. While Konschnik, et al. (1998) evaluated the use of administrative receipts in lieu of survey responses for small singleunit EINs, no previous study had evaluated a similar use of administrative end-of-year inventory data.

While comparing administrative end-of-year inventory to the reported 1999 ARTS end-of-year inventory, we observed that, for about half of the singleunit 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.

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.

For the rest of the paper, refer to the following definitions:

1) $$R_n^* = \frac{\Sigma_{i \in \Omega_n} w_i I_{i,n}}{\Sigma_{i \in \Omega_n} w_i S_{i,n}}$$

where i denotes a given singleunit EIN, n denotes a six-digit NAICS industry, w denotes the sampling weight assigned to i, I denotes end-of-year inventory as reported in the 1999 ARTS, S denotes sales as reported in the 1999 ARTS, Ω_n denotes all i classified in n with positive end-of-year inventory and sales reported in the 1999 ARTS.

2) $$R_{i,n} = \frac{I^A_{i,n}}{S^A_{i,n}}$$

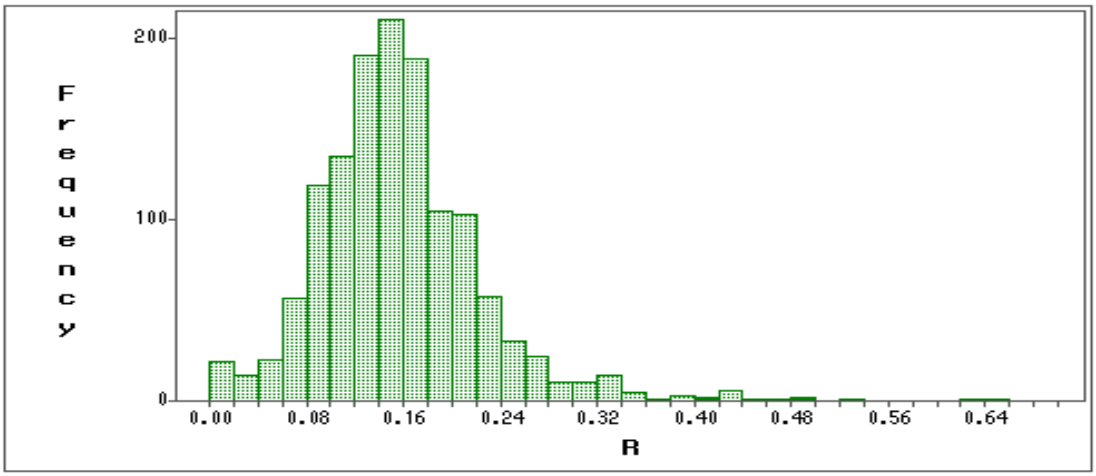
where $I^A_{i,n}$ = administrative end-of-year inventory for i classified in n, $S^A_{i,n}$ = administrative receipts for i classified in n, $S^A_{i,n} > 0$, and $I^A_{i,n} > 0$.

3) INV_FLAG = edit flag showing the reliability of administrative end-of-year inventory data.

To determine bounds on the value of $R_{i,n}$, we analyzed histograms of the ratio of reported inventory to reported sales for the seventy-one six-digit NAICS industries that had administrative inventory. The following histogram shows that this ratio ranged in value from 0.0014 to just above 0.65 for New Car Dealers (NAICS 441110).

Example of Finding Multipliers

For NAICS 441110



R* = 0.151630

To derive bounds for this industry, we noted that most of the reported inventory-to-sales ratios fell between the lower (0.0303) and upper (0.75815) bounds determined by multiplying R_n^* by 1/5 and 5, respectively. This result held true for other six-digit NAICS industries.

Listed below are steps we used to edit the administrative inventory data. For a given singleunit EIN, we would expect the ratio of administrative inventory to administrative receipts to fall between the lower and upper bounds on reported ratios for the EIN's six-digit NAICS industry. In Step 2, we divided the ratio by 100 because about half of the administrative inventory data was off by a factor of 100. The multipliers 1/3 and 3 were chosen for Step 2 so we could reduce the risk of having a legitimate outlier pass the edit.

Edit Steps: For i classified in n,

1) If $\frac{1}{5} R_n^* < R_{i,n} < \min(1,5 * R_n^*)$,

then INV_FLAG = 1.

2) Else, if $\frac{1}{3} R_n^* < \frac{R_{i,n}}{100} < \min(1,3 * R_n^*)$,

then INV_FLAG = 2.

3) Else, INV_FLAG = 0.

An analysis was done before and after the edit was performed, using reported 1999 ARTS end-of-year inventory for singleunit EINs. We examined singleunit EINs with both administrative and reported end-of-year inventory greater than zero to see how well the data for administrative end-of-year inventory resembled the reported value. There were 6,995 singleunit EINs where both administrative and reported end-of-year inventory were greater than zero. Of those 6,995 singleunit EINs, 946 EINs were assigned an INV_FLAG of 0; 3,499 EINs were assigned an INV_FLAG of 1; and 2,550 EINs were assigned an INV_FLAG of 2.

Before the edit was performed, the ratio of the sum of weighted administrative end-of-year inventory to the sum of weighted reported inventory for the 6,995 singleunit EINs was calculated to be 50.60. After the edit was performed, the same ratio was calculated to be 0.962. The following table shows the counts of six-digit NAICS industries by ranges of A, where A is defined as the ratio of the sum of weighted administrative end-of-year inventory to the sum of weighted reported inventory, before and after the administrative end-of-year inventory data were edited:

Table 2. Comparison of Administrative and 1999 ARTS Reported Inventory

A	**Counts of six-digit NAICS before the edit (% of total number of six-digit NAICS)**	**Counts of six-digit NAICS after the edit (% of total number of six-digit NAICS)**
0 < A <= 0.85	0 (0%)	5 (7.0%)
0.85 < A <=0.95	0 (0%)	15 (21.1%)
0.95 < A <= 1.05	0 (0%)	30 (42.3%)
1.05 < A <= 1.15	0 (0%)	12 (16.8%)
1.15 < A <=1.25	1 (1.4%)	7 (9.8%)
1.25 < A <=2	0 (0%)	1 (1.4%)
2 < A <=5	0 (0%)	1 (1.4%)
5 < A <=10	1 (1.4%)	0 (0%)
10 < A <=20	2 (2.8%)	0 (0%)
20 < A <=40	35 (49.3%)	0 (0%)
A > 40	32 (45.1%)	0 (0%)

After the edit, over 40% of the six-digit NAICS industries had a ratio between 0.95 and 1.05, and none of the ratios were greater than 5. The two six-digit NAICS industries, for which A > 1.25 after the edit, each had one singleunit EIN that passed the edit because both administrative receipts and administrative inventory were off by a factor of 100. Before the edit, only 6 of the 71 six-digit NAICS industries had reported inventory equal to administrative end-of-year inventory for at least 35% of the singleunit EINs classified. After the edit, 58 of the 71 six-digit NAICS industries fell into this scenario.

5. Effects of Incorporating Administrative Inventory Data in the 1999 Annual Retail Trade Survey

To analyze the uses of administrative inventory data for the 1999 ARTS, three imputation runs were done:

1) Using our current methodology, which produced what we call the *tabulated* estimates.

2) Substituting administrative inventory data, when available, or other imputed data for nonresponding singleunit EINs. This run will be discussed in this section.

3) Substituting administrative inventory data, when available, or other imputed data for small or nonresponding singleunit EINs. This run will be discussed in Section 6.

With the edit in place, we were able to use most of the administrative end-of-year inventory data for the imputation study. In our second run, to measure the effect of using the administrative end-of-year inventory, we added the following methods to the imputation methodology for ARTS:

1) If the EIN was a singleunit and not a death, with INV_FLAG = 1 and Administrative Inventory > 0, then Total Inventory = Administrative Inventory.

2) If the EIN was a singleunit and not a death, with INV_FLAG = 2 and Administrative Inventory > 0, then Total Inventory = Administrative Inventory/100.

Among the methods used to impute inventory data for ARTS, the two new methods were given a higher priority than the current methods. The estimate of total inventory for the Retail Trade (NAICS 44-45) increased only 0.41% from Run 1 to Run 2. Table 3 shows the percent contribution of reported, administrative, and other imputed data to the total inventory estimate for NAICS 44-45 from both imputation runs.

Table 3. Source of Total Inventory for NAICS 44-45 from Imputation Runs 1 and 2

	% Reported of Total Inventory	% Administrative Data of Total Inventory	% Other Imputed of Total Inventory
Imputation Run 1	86.86	0	13.14
Imputation Run 2	86.50	4.24	9.25

Table 4 displays the number of six-digit NAICS industries that have a percent change, within a specified range, between the total inventory estimates produced for Imputation Runs 1 and 2.

Table 4. Number of Industries by Percent Change in Total Inventory Estimate from Run 1 to 2

[-15%,-5%)	[-5%,-2%)	[-2%,-1%)	[-1%,0%)	[0%,1%)	[1%,2%)	[2%,5%)	[5%,15%)
0	1	5	17	22	8	13	5

For all 71 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.

6. Use of Administrative Receipts and Inventory Data for Small Singleunit EINs in the 1999 Annual Retail Trade Survey

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 singleunit EINs in the 1999 ARTS. Using the same methodology for identifying small singleunit EINs described in Section 3, a third imputation run was performed. In this run, reported sales and inventory data for small singleunit EINs were deleted and replaced with administrative data or other imputed data. Where available, administrative inventory data were used to impute data for small singleunit EINs and for singleunit EINs that did not respond to the 1999 ARTS. The estimate of total inventory for NAICS 44-45 increased only 0.09% from Run 1 to Run 3. Table 5 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 5. Source of Total Inventory for NAICS 44-45 from Imputation Runs 1, 2, and 3

	% Reported of Total Inventory	% Administrative Data of Total Inventory	% Other Imputed of Total Inventory
Imputation Run 1	86.86	0	13.14
Imputation Run 2	86.50	4.24	9.25
Imputation Run 3	77.70	10.94	11.37

Table 6 displays the number of six-digit NAICS industries that have a percent change, within a specified range, between the total inventory estimates produced for Imputation Runs 1 and 3.

Table 6. Number of Industries by Percent Change in Total Inventory Estimate from Run 1 to 3

[-15%,-5%)	[-5%,-2%)	[-2%,-1%)	[-1%,0%)	[0%,1%)	[1%,2%)	[2%,5%)	[5%,15%)
2	7	5	14	10	12	15	6

Though the estimates from Imputation Run 3 seem to look good, for some six-digit NAICS industries, the estimated CV of the total inventory estimate from Run 3 was substantially larger than the one from Run 1. Further investigation is required to determine what may be causing the high estimated CVs and if administrative end-of-year inventory can be used for small singleunit EINs classified in particular NAICS industries.

7. Continuing Research

In addition to completing the inventory research, we plan to study the effects of not mailing annual forms to additional singleunit EINs on 1999 and 2000 ARTS, ATS, and SAS estimates at six-digit NAICS levels. The estimates evaluated will include those for which administrative receipts and end-of-year inventory data would be used to impute values for all nonmailed singleunit EINs. The results of this study will be used to determine whether to withhold additional singleunit EINs from the 2001 mailings.

In addition to the work described in the sections 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.
- compare administrative expenses to expenses reported from survey forms.
- 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

Kinyon, D., D. Glassbrenner, J. Black, and R. Detlefsen (2000), "Designing Business Samples Used for Surveys Conducted by the United States Bureau of the Census," paper presented at the second International Conference on Establishment Surveys, Buffalo, NY.

Konschnik, C., J. Johnson, and J. Burton (1998), "The Use of Administrative Records in Current Business Surveys and Censuses," report of the NAICS Frame and Sampling Issues Working Group.

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