THE USE OF ADMINISTRATIVE RECORDS IN CURRENT BUSINESS SURVEYS AND CENSUSES

Carl A. Konschnik, Jennifer S. Johnson and James N. Burton, Bureau of the Census*
Carl A. Konschnik, Bureau of the Census, SSSD, Washington, DC 20233

Key words: Retail sales, inventory, payroll, business tax returns, editing and imputation.

1. Introduction

Administrative records play a central role in the Census Bureau's production of economic statistics. To give a few examples, they help us: construct frames for surveys and censuses; identify new businesses (births) and alert.us to firms that have gone out of business (deaths); impute for nonresponse to a survey inquiry; reduce respondent burden and costs by using administrative data for census tabulations rather than mailing and processing a census form; and, improve measures of size for sampling. Although administrative records are used extensively to produce statistics in virtually all economic areas as well as in the demographic area, this paper will discuss their role chiefly in those current business surveys and censuses that measure a large portion of the services sector of the U.S. economy. Specifically, this portion includes retail and wholesale trade, and the many components of services industries, such as personal services, medical and legal services, transportation, communications, and others. Note that although our full research paper essentially included results for all these areas, we have had to limit ourselves to showing results only for the retail trade in this Proceedings version.

On a regular basis the Census Bureau receives administrative record data from other Federal agencies, chiefly from the Internal Revenue Service (IRS), and also to a lesser extent from the Social Security Administration (SSA), and the Bureau of Labor Statistics (BLS). These data form the basis for the construction and maintenance of the Census Bureau's central business register, the Standard Statistical Establishment List (SSEL). The SSEL contains data on all known employer business establishments in the U.S. Here we define business establishment as a specific physical location or structure where business is conducted. By way of example, stores, shops, offices, and manufacturing plants are examples of business establishments. Sidewalk vendors' carts or wagons are not considered business establishments; instead we consider their home base as the establishment. Key administrative record data on the SSEL are business or owner's name and address, legal form of ownership, industrial classification, quarterly and annual payroll, number of employees, annual sales or receipts, and company affiliation. Using this and other information, coupled with statistical data collected by the Census Bureau, we are able to efficiently and cost effectively operate our statistical programs.

Even though we have used administrative record data in many ways for a long time, we continue to search for ways to improve our statistical programs in terms of reduced burden on the business community and more efficient use of resources, while at the same time expanding the coverage and quality of our statistical products. Thus, the primary focus of our recent research in the administrative record area has been to identify and assess the impact of expanded uses of administrative records for our programs. More specifically, we have been looking for additional ways to use administrative record data in lieu of mailing a report form in our annual surveys of retail and wholesale trades, as well as those in the services industries. This paper focuses on our investigation and evaluation of this increased use of administrative record data. This work becomes more important now as we garner our resources for introducing the new North American Industry Classification System (NAICS) to replace the existing Standard Industrial Classification (SIC) system over the next few years. Indeed, the 1997 Economic Census will be published on a NAICS basis, and the current business surveys, defined here as the monthly surveys in retail and wholesale, and the annual surveys in retail, wholesale, and services, will start publishing on a NAICS basis in early 2001.

2. Existing Uses of Administrative Records

All employer businesses must use the Federal Employer Identification Number (EIN) as their primary taxpayer identification number when reporting payroll data to the IRS. They do this in each quarter of the year using IRS Form 941, "Employer's Quarterly Federal Tax Return". In addition, if their legal form of organization is corporation or partnership, they also must file their annual income tax returns using their EIN. Sole proprietorship firms, however, must file their annual income tax return using the owner's Social Security Number (SSN). The IRS regularly transfers selected data from these tax returns to the Census Bureau for the specific and sole use of producing official statistics for the U.S., in accordance with the requirements and safeguards provided for by public law. As a result, the SSEL uses the EIN as its primary identifier for single establishment business firms (singleunits) and as a secondary identifier for multi-establishment (multiunit) firms. The SSEL's primary identifier for multiunit firms is a Census Bureau assigned six-digit company number, called an alpha number. Each individual establishment of a multiunit firm or company has as its primary identifier the company's six-digit alpha number followed by a four-digit plant or establishment number within the company. Because every multiunit employer establishment must have some EIN used for tax reporting purposes, this EIN becomes its secondary identifier. An EIN may be used for only one establishment of a firm, or it may be used for several or all establishments of that firm. Situations where establishments share an EIN can create some difficulty in assigning the correct percentage of an EIN's administrative record data to each establishment. This does not necessarily pose a problem in surveys where the desired reporting unit is the EIN level record and all establishments associated with that EIN share the same kind of business activity. This is probably the case for the majority of multiunit EIN reporting units in the retail, wholesale, and services surveys.

Here are some of the major ways in which we use administrative records, reported at the EIN or SSN level:

- Administrative records provide the basis for building statistical frames of business units that we use for our economic census and all major business surveys.

- We use administrative payroll and sales (and to a much lesser extent employment) data to help in determining measures of size for sampling.

- We use administrative payroll, sales, and employment data for editing survey reports and imputing for nonresponse.

- Survey births and deaths are identified and the frames and samples are updated and maintained. This process allows us to control each survey and ensure that the sample properly represents the universe at all times.

- For our economic censuses we use administrative sales, payroll, and employment data in lieu of mailing a census report form to singleunit establishments with payroll below certain levels. These payroll levels are determined for prescribed SIC levels so that about 80% of the dollar volume of payroll comes from multiunit and singleunit establishments mailed a census form. Administrative record sales are therefore used for a large number of singleunits whose sales aggregate to probably somewhat less than 20% of the SIC's total dollar volume.

- Administrative record data provide kind of business classification information. We obtain business classification that is reported on tax return forms from the IRS. These include Form 1040, Schedule C for sole proprietorships, Form 1120 for corporations, and Form 1065 for partnership businesses. From SSA we receive SICs coded based on IRS Form SS-4, "Application for an Employer Identification Number". We also receive classification data from BLS, generally on a quarterly basis, that gives us SIC codes for units that we have not had sufficient (or, possibly, any) information on which to base a code.

- For our censuses and annual surveys of retail and services, we use Form 1040, Schedule C tax return data to define and measure the sales or receipts and number of establishments of nonemployer businesses. If a sole proprietorship business has employees, they are asked to include their EIN on their tax return. This enables us to distinguish employers filing Form 1040, Schedule C from nonemployers filing the same type of tax form.

Without administrative record data, the cost of producing reliable statistical estimates of the Nation's economy would be prohibitive. With them, we are able to efficiently and cost effectively operate major statistical programs that are of high quality. The following section discusses how we might make more use of these important data sources.

3. Research into Expanded Uses of Administrative Records

We have been looking at the impact of replacing sales or receipts data collected via survey forms with annual tax return data for singleunit establishments in our annual surveys of retail, wholesale and services. We concentrated on singleunits because the EIN level administrative record sales required no disaggregation and thus provide reliable measures of the (singleunit) firm's sales. Also, restricting administrative source data to singleunits in lieu of mailing a questionnaire is consistent with our economic census procedures. However, because in our surveys we are collecting firm or EIN level data and not establishment level data, we also plan to later look at using administrative records for EIN multiunits in cases where all associated establishments have the same SIC.

We begin with a discussion of our research for the sales variable for the Annual Retail Trade Survey (ARTS). After that, we also give the results of our evaluation of the other key variables in the ARTS of potential non mailed units (units for which we will tabulate sales based on their administrative sales data, if these data are available) using data imputed based on sales. These variables are end-of-year inventory, annual purchases, and accounts receivable balances. We currently receive no administrative data for these variables. However, beginning in early 1999 IRS will start sending us end-of-year inventory data as reported on the 1998 business income tax returns. High cost prevents us from getting purchases data from tax returns, and accounts receivable data are not reported on tax returns.

3.1 Retail Sales

We present here Tables 1a., 1b., and 1c. that show the contribution of the noncertainty (sample weight >1) singleunit EINs to the 1996 total ARTS sales estimates at the total retail level. There were 26,297 sample units with tabulated sales greater than 0 for the 1996 estimates. This includes units of all types, namely, certainty (sample weight = 1, or self-representing) alpha companies, noncertainty multiunit EINs, certainty singleunit EINs, and noncertainty singleunit EINs. The percentages in the tables are relative to this overall number of units shown in parentheses. The dollar volume percentages are relative to the 1996 ARTS total dollar value as calculated for each of the specific tables. This total is also shown in parentheses in each table. In terms of dollar volume, the noncertainty singleunit EINs made up about 40%. This compares with alpha companies that made up about 44%, noncertainty multiunit EINs about 12.5%, certainty singleunit EINs about 1%, and nonemployers about 2.5%.

Table 1a. shows the 1996 ARTS as tabulated. For noncertainty singleunit (SU) EINs the approximate number of units mailed was 20,123. These included units that were active in 1995 but inactive in 1996. Because we collected data for both years in the survey, they had to be included in the mailing. Therefore many were tabulated at a nonzero sales value for 1995 but zero for 1996. 17,955 noncertainty SU EIN units had nonzero sales tabulated for 1996. Note that the noncertainty SU reporters contributed 33% of the total dollar volume of the 1996 ARTS sales estimate. For noncertainty SUs that did not report we used administrative sales data where available, and imputed as a final option. Administrative sales data contributed 5% and imputed data 2%.

Table 1a.

Noncertainty SU EIN Contribution to
1996 Total Retail Sales
(As Tabulated With Sales > 0

	Total	Reported	Imputed	Admin Sales
No. of Noncert. SU EINs	17,955	12,606	2,088	3,261
% of Units of All Types (26,297)	68%	48%	8%	12%
Noncertainty SU EIN Sales (in billions)	$973	$804	$51	$118
% of Total Retail Sales ($2,460 billion)	40%	33%	2%	5%

We did our imputation in the following priority order:

1. If a unit had nonzero previous year sales (PS), nonzero previous year payroll (PP) and current year payroll (CP), then current year sales (CS) was defined as:

$$CS = PS \times \frac{CP}{PP} .$$

2. If a unit was in the monthly survey and had more than 9 months of monthly data reported, then CS was set to the sum of the monthly data for all 12 months.

3. CS = B x CP, where B is a regression derived sales to payroll factor determined at the (essentially four-digit) SIC level.

Table 1b. shows the effect of potentially not mailing any of the noncertainty SU EINs. Proceeding with this scenario, we would withhold from mailing 20,123 singleunits. This represents the 17,955 tabbed at nonzero plus an additional 2,168 mailed but tabbed at 0 for 1996 but at a nonzero value for the prior year (1995). The total retail estimate would increase to $2,472 billion. This is an increase of 0.48% over the 1996 ARTS as tabbed. Of the noncertainty SU reporters' contribution of 33% of the dollar volume of total ARTS sales as seen from Table 1a., 29% is replaced by administrative data (administrative sales data share increased from 5% to 34%) and the remaining 4% is replaced by imputation. For this table and for Table 1c. which follows, we made one modification in the imputation scheme above. Essentially, we used the priority 1. imputation scheme only if the unit had all four quarters of nonzero payroll for the current year and prior year. This made a significant difference for a relatively few prominent cases.

Table 1b.

Noncertainty SU EIN Contribution to
1996 Total Retail Sales
Mailing No Noncertainty SU EINs
(As Potentially Tabulated With Sales > 0

	Total	Reported	Imputed	Admin Sales
No. of Noncert. SU EINs	17,955	0	4,169	13,786
% of Units of All Types (26,297)	68%	0%	16%	52%
Noncertainty SU EIN Sales (in billions)	$985	$0	$148	$837
% of Total Retail Sales ($2,472 billion)	40%	0%	6%	34%

At the four-digit SIC levels, the percentage differences are larger, with the largest at about 17%, 19 kinds of business having differences that exceed 2%, and 32 kinds of business having differences that exceed 1%. For our research, we produced tables that show these differences at all the four-digit retail SIC levels. These tables are obviously too extensive for these proceedings, but they give us important information on which to base our future plans.

Table 1c. shows the anticipated effect of potentially mailing only those noncertainty SU EINs with payroll exceeding the payroll cutoff limits used for the 1997 Economic Census. Essentially, we would mail the units with payroll above cutoffs, but use administrative records or imputation for those with payroll below these cutoffs. For Table 1c., the mailing count would be 7,674 as compared with 20,123 noncertainty SU EIN units mailed. This represents a savings of 12, 449 units in terms of response burden and processing costs. Note also that Table 1c. does not show the 7,674 count as mailed because only 6,050 actually reported. The remaining 1624 were accounted for by imputation or administrative data. It is noteworthy that these savings can potentially be achieved without a change in the total dollar volume estimate for retail sales. Indeed the estimates are exact to the nearest billion dollars. The actual difference is about 0.01%. Under this scenario, the noncertainty SU EINs mailed retain 23% of the total dollar volume (as compared to 33% as shown in Table 1a.) and the remaining 10% is split, 8% to administrative records and 2% to imputation.

Table 1c.

Noncertainty SU EIN Contribution to
1996 Total Retail Sales
Mailing Only Noncertainty SU EINs With Payroll
Above Census Cutoffs
(As Potentially Tabulated With Sales > 0

	Total	Reported	Imputed	Admin Sales
No. of Noncert. SU EINs	17,955	6,050	3,382	8,523
% of Units of All Types (26,297)	68%	23%	13%	32%
Noncertainty SU EIN Sales (in billions)	$973	$558	$90	$325
% of Total Retail Sales ($2,460 billion)	40%	23%	4%	13%

We also produced extensive tables that showed that sales differences at various SIC levels are larger than for total retail sales, with the highest at 4.6%, and only 12 four-digit retail SICs with a difference exceeding 1%. None of these are for any major publishable levels. These are relatively small differences for these levels.

3.2 **Retail Inventory**

By inventory here we mean the non-LIFO (Last In First Out) or pre-LIFO inventory valuation. Thus, if a business uses the LIFO valuation method, the LIFO reserve is (assumed to be) added to the LIFO value. Mailing none of the 17,955 noncertainty SU EINs described in the previous section and imputing for end-of-year inventory for these units would yield a total retail inventory estimate of $319 billion. This is 2.47% higher than the $312 billion estimate tabulated for in the 1996 ARTS. This compares with the 0.48% difference observed for sales under the same scenario. By way of comparison, if we were to mail the 7,674 units with payroll above the cutoffs, then the inventory estimate would be only 0.6% higher than those produced for the 1996 ARTS ($314 billion versus $312 billion).

For units potentially not mailed or for nonresponding units, the imputation of inventory was done essentially as follows. First, if prior year sales (PS) and prior year inventory (IP) were both nonzero, and denoting current year sales by CS, then the current year inventory (CI) was set to

$$CI = CS \times \frac{IP}{PS} .$$

If these conditions were not met, then CI = CS x IR, where IR is an inventory to sales ratio developed based on reporters at the SIC level.

3.3 Retail Purchases

Retail purchases data are used to produce cost of goods sold estimates, which are in turn used to determine gross margin (GM) estimates for retail sales. Specifically, denoting retail sales by S, cost of goods sold by CG, beginning of year inventory as BI, end-of-year inventory as EI and purchases as P, then CG = BI + P- EI. Gross margin is then defined as: GM = S - CG. Mailing none of the 17,955 noncertainty SU EINs and imputing for purchases would give a total retail purchases estimate of $1,698 billion. This is just 0.37% higher than the estimate of $1,692 billion obtained from the 1996 ARTS. Further, if we were to mail noncertainty SU EINs with payroll above the census cutoffs, the purchases estimates would differ by less than a billion dollars, or .04%. We used the same imputation scheme as described in 3.2 above with purchases replacing inventory.

3.4 Retail Accounts Receivable

The initial results we observed for total retail accounts receivable balances is equally encouraging. There is a 1.12% difference between the estimate obtained from imputing for the (potentially) nonmailed noncertainty SU EINs ($63.33 billion) and the 1996 ARTS estimate ($62.63 billion). If we were to mail the large payroll units (7,674) the estimates would differ by only 0.22% (62.77 billion versus 62.63 billion). The method of imputation we used for these tables is the same as that which we used for inventory, with the appropriate replacement of the inventory variable with the accounts receivable variable. One serious problem here is that not all retail EINs carry accounts recievable balances type of credit. We will look at various options to address this problem.

3.5 Percentage of Units With Administrative Sales

An important measure in the decision to use administrative records to substitute for reported annual sales data is the percentage of sample units that have administrative sales data on the SSEL. As of the first week of December 1997, (the approximate time at which these data would be used in production) 77% of the singleunit EINs with nonzero payroll in the ARTS sample had current year (1996) administrative sales. Eighty-three percent had previous year (1995) administrative sales. There are several factors that cause these deficiencies. One common reason is our inability to link all EIN based payroll records with their associated SSN based sales records for sole proprietorship businesses. Another is the reporting of payroll under one EIN and sales under another which also complicates our attempts to link payroll and sales data for a specific EIN.

3.6 Agreement Between Administrative Sales and Reported Sales

Although using administrative sales data as a replacement for reported survey data appears to be a viable approach when looking at total retail or even SIC totals, there are nonetheless differences for specific cases. For example, for the 10,402 noncertainty SU EINs with nonzero payroll in retail that also have nonzero, current year data reported on the 1996 ARTS form, 42% have the same (but for minor rounding) value, 57% are within 1%, 75% are within 5%, and 85% are within 10%. For the remaining 15% of the cases, 4% have reported sales greater than administrative sales, and 11% have administrative sales greater than reported sales. We also computed the overall ratio of weighted administrative sales and weighted reported sales for the noncertainty SU EINs that had nonzero values for both of these items. This ratio was 1.018, indicating that replacing reported sales with administrative sales for these noncertainty SU EINSs would increase the sales estimate by 1.8%. Nonetheless, from Section 3.1 and Table 1b., we saw that using administrative data for all the noncertainty SU EINs would increase the total ARTS estimate by only about 0.48%. This is so, of course, because the noncertainty SU EINs comprise only about 40% of the total sales estimate, and not all noncertainty SU EINs had both administrative and reported sales.

There are several reasons why administrative sales might be higher or lower than reported sales. Some major reasons are given here.

- The administrative data are on fiscal year basis rather than the calendar year basis requested on the survey reports.

- The administrative data for the EINs may include nonretail sales or receipts, either because of nonretail activities of the single establishment or because of possibly additional establishments that are not in retail trade and for which we are unaware of their existence.

- There may be errors in either the administrative data or the reported data. These could be due to errors in completing the respective forms, keying errors, or other processing errors.

- Some business firms use one EIN for quarterly payroll reporting but a different EIN for filing their annual income tax. Thus, the tax return data could contain the sales for several payroll EINs.

3.7 Additional Paths of Research

We have begun to explore specific problems and options associated with using administrative record data to best advantage. Some of these are given in the tasks and options listed below.

- Studying further the availability and degree of agreement between administrative sales and reported sales for specific units and understanding the reasons for these differences.

- Adjusting administrative sales data prior to them using in tabulations, using the ratio of survey reported data to tax return sales data.

- Using administrative sales data for some or all noncertainty multiunit EINs.

- Selecting a trace sample of noncertainty SUs for mailing. This would include units above and below payroll cutoffs, or, alternatively, include only units below cutoffs. This trace sample would be used to develop relationships for imputing nonmailed units.

- Determining appropriate editing procedures for administrative data prior to using them for survey tabulations.

- Identifying specific sample units for which administrative data would be expected to agree with data reported on a survey form. Directly related to this is the task of identifying specific units for which administrative data are not an appropriate alternative to survey response data.

4. Conclusions and Future Plans

It is fairly clear that using administrative data from tax returns will be a viable alternative to mailing large numbers singleunit sample units and possible also some specific classes of multiunits. We are continuing research to precisely target these sample units. Once our investigation is completed, we will be making recommendations to program managers of the various surveys that detail specific steps to consider. From what we've seen so far, we can reasonably expect to significantly reduce processing costs and respondent burden by expanding the use of administrative records.

References

Papers and reports on administrative records use in the Federal statistical system are numerous. Some of the specific documents that have influenced the perspective for the work in our current paper are listed below.

Office of Federal Statistical Policy and Standards (1980), Report on Statistical Uses of Administrative Records. Statistical Policy Working Paper 6.

U.S. Department of the Treasury, Internal Revenue Service (1982), Statistics of Income and Related Administrative Record Research: 1982.

U.S. Department of the Treasury. Internal Revenue Service (1984), Statistical Uses of Administrative Records: Recent Research and Present Prospects, Volumes I and II.

* This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.