The Use of Administrative Data to Compute Measure of Size for New Employer Births[1]

Lora M. Gillott, James N. Burton, and Carol S. King
U.S. Census Bureau
Lora.M.Gillott@census.gov

Abstract

The United States Census Bureau samples new employer births quarterly using a two-phase sample design. The source of these new employer births is the Census Bureau's Business Register. The Business Register is a multi-relational database that contains a record for each known establishment in the United States and is periodically updated with administrative data from a variety of sources, including the Internal Revenue Service and Social Security Administration.

In the first phase of the sample design, new employer births are identified. A sample of these is mailed a classification form. In addition to providing new or refined industry classification codes, the completed form provides sales, receipts or revenue, company organization, and other key information needed for sampling. If a response is available, the reported two months of sales (receipts) may be used to compute the measure of size for second-phase sampling. However, not all businesses respond, so it is necessary to use administrative payroll data to compute the measure of size for these.

This paper documents the research that looks at using administrative payroll data to calculate the measure of size for all new employer births. This research is a continuation of research to determine the feasibility of eliminating the first phase of the new employer birth sample selection, basing the measure of size only on administrative data.

Keywords: administrative data; two-phase sampling; measure of size

1. Introduction

Each quarter, the United States Census Bureau receives administrative payroll records for new employer births from the Internal Revenue Service (IRS). These new employer births are added to the Census Bureau's Business Register. The Business Register serves as the source of the frame for the Census Bureau's monthly and annual surveys of the retail, wholesale, and service sectors. For more details about the Business Register, see Walker (1997) and Konschnik and Walker (1999). For this research, only the annual surveys were analyzed: Annual Retail Trade Survey (ARTS), Annual Trade Survey (ATS), and Service Annual Survey (SAS). SAS is further separated into Computers (SAS-C), Finance (SAS-F), General (SAS-G), Health (SAS-H), Information (SAS-I), and Transportation (SAS-T).

Currently, for each of these surveys, a two-phase sample design is used to select new employer births identified each quarter from the Business Register. Burton, King, and Hunt (2001) present reasons for the current two-phase design, as well as improvements to industry coding performed by the Social Security Administration that have motivated the reevaluation of the current sample design. Using the two-phase design, new employer births are added to the Census Bureau's surveys approximately nine months after they begin operation. Eliminating the first phase of the design would allow new employer births to be added to the surveys approximately three months earlier and eliminate the cost of first-phase mailing and data collection operations.

In the first phase of the design, a sample of new employer births is mailed the U.S. Department of Commerce's SQ-CLASS form. Responses from this form help with industry classification and calculation of a measure of size (MOS) for use in second-phase stratification. Industries are classified using the 1997 North American Industry Classification System (NAICS). NAICS is a six-digit detailed industry classification system in which the first two digits represent the sectors agreed upon by Canada, Mexico, and the United States. For more information on NAICS classification, see U.S. Office of

[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, Michael Kornbau, William C. Davie Jr., and David L. Kinyon for their helpful comments.

Management and Budget (1998). Incorrect industry classification creates misrepresentation of new employer births in the Census Bureau's surveys because improper weights are assigned to the selected new employer births.

The first step in the second phase of the sampling process is to determine if a new employer birth meets the criteria for sampling. A MOS is then calculated for each eligible new employer birth. It is then determined if the eligible new employer births are already represented in the surveys. If so, they are dropped from further consideration. A second-phase sampling weight is assigned to each of the remaining units based on the strata in which the new employer births are assigned. First-phase and second-phase sampling weights are then compared to determine which units to subject to second-phase sampling. New employer births whose first-phase sampling weight is greater than or equal to twice the second-phase sampling weight are selected into the second-phase sample and assigned their first-phase weight. The rest are subjected to second-phase non-certainty sampling using stratified systematic sampling.

To select a representative sample of new employer births, two items are needed: accurate industry classification and an accurate MOS. For more information on accurate industry classification, see Burton, King, and Hunt (2001). Each new employer birth's estimated annual total sales are calculated to produce a MOS. The estimated annual total sales are calculated from responses to the SQ-CLASS form with current month and/or prior month sales or from administrative payroll data. New employer births are selected in the retail, service, merchant wholesale, and nonmerchant wholesale sectors.

Sections 2, 3, and 4 discuss three different methods used to compare the monthly sales MOS and the administrative payroll MOS. In Section 2, new employer births are assigned to strata using two size measures: a MOS based on administrative payroll data and a MOS based on reported monthly sales data. For each new employer birth, the two strata are compared to determine the differences and effects on sampling of each. Large differences between strata are of concern, because the weight assigned to a new employer birth would be very different, depending on which method was used. In Section 3, a new sample is selected using only payroll to determine the MOS. The new sample is compared to the sample drawn using current methodology to determine how the estimates and variances will be affected by using only administrative payroll data to calculate a MOS. In Section 4, new employer births that reported to the 2001 annual surveys are stratified based on their reported sales, and these strata are compared to the ones determined using the administrative payroll MOS and the monthly sales MOS. The reported annual sales data is the actual MOS for the employer birth; therefore, the estimate that produces the MOS closest to the true MOS is more accurate. Finally, Section 5 summarizes our findings and suggests further research.

2. Comparing Two Size Measures: Administrative Payroll MOS and Monthly Sales MOS

For this study, new employer births identified during the period from fourth quarter 2001 to third quarter 2002 were analyzed. This period corresponds to the period in which new employer births were added to the 2001 annual surveys. For each 2001 annual survey, Chart 1 shows the number of new employer births using administrative payroll data to determine the MOS and the number of new employer births using monthly sales data to determine the MOS. Using the current methodology, the MOS was calculated using administrative payroll data for about half of the new employer births.

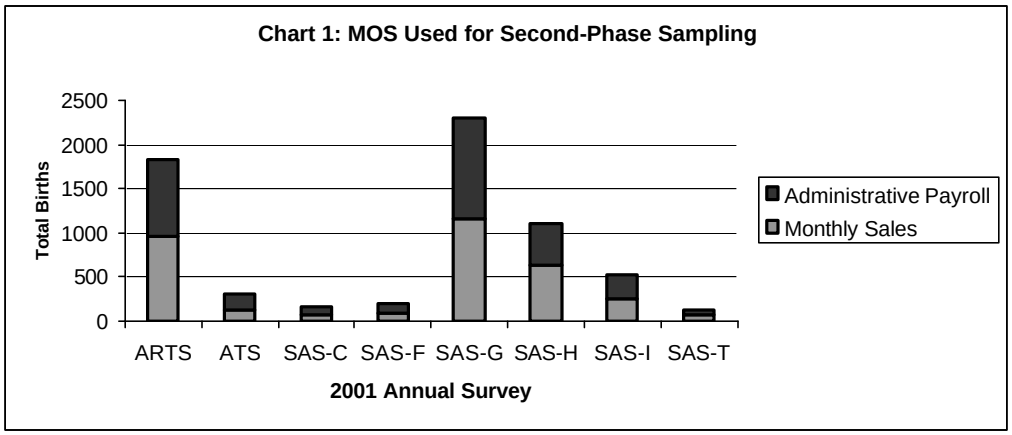

Table 1 compares the differences between the stratum assigned to a new employer birth using the monthly sales data MOS and the stratum assigned using the administrative payroll MOS. Only new employer births having both reported data from

the SQ-CLASS form and administrative payroll data were included in this analysis. Strata were constructed by industry based on a MOS that approximates annual sales on a 1997 basis, with the maximum number of strata being thirteen. The strata for each industry were then numbered in sequential order, starting with the smallest stratum. For more details about how the strata were constructed see Kinyon, Glassbrenner, Black, and Detlefsen (2000). The difference between the two strata was calculated as the stratum determined using the monthly sales MOS minus the stratum determined using the administrative payroll MOS. If the difference between the two stratum classifications is small, this shows there is little difference between the two MOSs for a given new employer birth. However, a large difference causes concern because this would mean the two MOSs are very different, given that the width of the strata increases, as the strata get larger. This would also mean the weight assigned to the new employer birth would be very different, depending on which method is used.

Some key points from Table 1:

- 36.07% of the fourth quarter 2001 through third quarter 2002 new employer births subjected to second-phase sampling had the same stratum assigned, regardless of how the MOS was determined.
  - ✓ 41.25% for merchant wholesale
  - ✓ 35.62% for retail
  - ✓ 35.85% for service
- 71.24% of the new employer births were determined to be within one stratum (equal or (+/-) one stratum difference) using the two MOS methods.
  - ✓ 74.76% for merchant wholesale
  - ✓ 74.23% for retail
  - ✓ 69.21% for service

Table 1: Percent Distributions of New Employer Births by Difference between Stratum Determined Using Monthly Sales and Stratum Determined Using Payroll

	Stratum Differences												
Sector	Equal	1	-1	2	-2	3	-3	4	-4	5	-5	6 to 12	-6 to -12
Merchant Wholesale	41.25	11.61	21.90	3.61	9.32	1.85	3.96	0.97	1.32	0.53	1.05	1.22	1.41
Retail	35.62	12.02	26.59	4.36	10.41	1.35	3.82	0.79	1.96	0.46	0.99	0.55	1.08
Service	35.85	17.34	16.02	7.09	7.27	3.09	3.38	1.85	2.06	1.01	1.29	1.42	2.33
Total	36.07	15.20	19.97	5.96	8.46	2.43	3.56	1.44	1.98	0.80	1.17	1.11	1.85

Of the new employer births, 37.00% had a larger stratum determined when payroll was used to compute the MOS, whereas 26.93% of the new employer births had a larger stratum determined when monthly sales was used to compute the MOS. Table 2 shows a breakout of these differences by sector. For the retail and merchant wholesale sectors, the strata determined using administrative payroll were larger approximately twice as often as the monthly sales strata. For the service sector, the two MOSs produced larger strata at approximately the same rate. A larger stratum allows a new employer birth to be assigned a more conservative weight. This means that a birth will be sampled at a higher rate, which is desirable because it is difficult to predict a new employer birth's growth over time. Therefore, the most conservative weight should be applied to a new employer birth.

Table 2: MOS Stratum Comparisons

Sector	Larger Stratum from Administrative Payroll	Larger Stratum from Monthly Sales	Same Stratum
Merchant Wholesale	38.96	19.79	41.25
Retail	44.84	19.54	35.62
Service	32.34	31.81	35.85
Total	37.00	26.93	36.07

3. Comparing Current Methodology to Administrative Data Methodology

To determine how using only the administrative payroll data MOS would compare to the current method, a new second-phase sample was selected, based on results reported by the first-phase sample, using only administrative payroll data to determine the MOS. For the remainder of this paper, the new sample will be called s_2 and the original sample will be called s_1. S_1 contained 7,485 new employer births and s_2 contained 6,889 new employer births. Due to the overlap in the methodology and because both samples were selected independently, 4,687 new employer births were the same in both samples. So that the results of the two samples could be compared, reported data from s_1 was not used, instead sales or revenue for new employer births from both samples were imputed. New employer births selected in s_2 would not have the

chance to report, as did the new employer births in s_1; therefore, if the data for the new employer births in s_1 were not imputed, the two samples would not be comparable.

For ARTS and ATS, sales for the survey year are collected. For SAS, revenue data are collected. Throughout this paper, however, both will be called sales. For the businesses that were not new employer births represented in ARTS, ATS, and SAS, reported data, when available, were used to determine the estimates; otherwise, sales were imputed. The following hierarchy was used to impute sales for new employer births:

1. If current-year administrative payroll equals 0 then sales equals 0.
2. If administrative sales are greater then zero and are between 1/3*β*current-year administrative payroll and 3*β*current-year administrative payroll, then sales equals administrative sales. β is the estimated slope of a zero-intercept resistant regression model of sales to payroll for establishments in the 1997 Economic Census at detailed NAICS levels.
3. If current-year administrative payroll is greater than zero, then sales equals β*administrative payroll.
4. Else, sales equals zero.

Sales for new employer births with current-year administrative payroll equal to zero, but administrative sales greater than zero, are imputed as zero because these new employer births are assumed to be nonemployers, or businesses with no paid employees. These new employer births are represented in the United States Census Bureau's nonemployer statistics.

Table 3: Sample Estimate Comparisons

			Total Estimates		**Standard Error Estimates**	
SURVEY	**NAICS**	**NAICS Description**	**Complete Sample Percent Difference**	**New Employer Birth Sample Percent Difference**	**Complete Sample se (x_2) / se (x_1)**	**New Employer Birth Sample se (x_2) / se (x_1)**
SAS-C	54XXXX	Professional, Scientific, and Technical Services	-0.1537	-4.7761	1.0492	1.6168
SAS-F	52XXXX	Finance and Insurance	-0.4575	-16.9607	1.0465	0.6218
SAS-G	53XXXX	Real Estate and Rental and Leasing	0.3620	8.7201	1.0087	0.8600
	54XXXX	Professional, Scientific, and Technical Services	0.3097	8.7247	1.0370	1.2822
	56XXXX	Administrative Support and Waste Management and Remediation Services	0.3621	6.5367	0.9631	0.6040
	71XXXX	Arts, Entertainment, and Recreation	0.0945	3.8078	1.0049	0.6235
	81XXXX	Other Services Except Public Administration	0.7207	41.4812	1.0195	1.6613
SAS-H	**62XXXX**	**Health Care and Social Assistance**	**0.3128**	**20.4364**	**1.0829**	**3.1478**
SAS-I	51XXXX	Information	0.0701	3.4402	0.9931	0.5852
SAS-T	48XXXX	Transportation and Warehousing	-0.2573	-13.5891	1.0379	0.4438
	49XXXX	Transportation and Warehousing	-0.1502	-10.7840	0.9803	0.1908
ATS	42XXXX	Wholesale Trade	0.1564	5.5709	0.9997	0.9265
ARTS	**44XXXX**	**Retail Trade**	**-1.9206**	**-3.8634**	**1.1035**	**1.6463**
	45XXXX	Retail Trade	0.4345	52.3967	1.0221	0.5783
	72XXXX	Accommodation and Food Services	0.9424	5.0962	0.9732	0.9890

Table 3 shows the percent difference between the total sales estimates, x_1 and x_2, for both the overall total and the new employer birth sample by the 2-digit NAICS for ARTS, ATS, and SAS. Because samples are selected every five years, after the Economic Census is completed, the effects new employer births have on the annual estimates will grow through the life of the sample. This study was conducted approximately halfway through the current sample. Percent differences are

computed as $[(x_2 - x_1)/x_1]*100$, where x_i, i = 1 or 2, is the estimate using sample one or two. The complete sample percent differences were less than one percent for all industries except NAICS 44XXXX.

The ratios of the estimated standard error of x_2, se (x_2), divided by the standard error of x_1, se (x_1), were also compared to determine the efficiency of using s_2. Estimated standard errors of overall estimates for NAICS codes other than 44XXXX and 62XXXX appeared not to change much by computing the MOS from only administrative data. However, NAICS 44XXXX appeared to have a 10.35% increase in estimated standard error using the administrative payroll MOS, and NAICS 62XXXX appeared to have an 8.29% increase in estimated standard error.

4. Comparing Response Data from the 2001 Annual Surveys to the Payroll MOS and the Monthly Sales MOS

There were 7,485 new employer births selected using current methodology to determine the MOS. The percent of new employer births that reported to the 2001 annual surveys were 59.1 % for ARTS, 54.0 % for ATS, 62.2 % for SAS-C, 55.9 % for SAS-F, 61.4 % for SAS-G, 64.4 % for SAS-H, 54.1 % for SAS-I, and 68.5 % for SAS-T. The new employer births that reported to the 2001 annual surveys that had administrative payroll and monthly sales data available from the SQ-CLASS form were analyzed further to determine which MOS was more accurate. The annual sales data reported in a 2001 annual survey was assumed to be the true MOS for the new employer birth; therefore, the estimate that produced the MOS closest to the true MOS was considered the most accurate.

For each of these new employer births, a MOS based on the annual sales reported in a 2001 annual survey was constructed and the appropriate numbered stratum was assigned. The MOS was constructed by annualizing the response to the survey by an appropriate factor based on the number of quarters the new employer birth was in business. The factor applied was determined by dividing four by the number of quarters the employer birth was in business. This value was then multiplied by a deflation factor to convert the MOS to a 1997 level (Recall that the sampling strata were constructed based on a MOS that approximated annual sales on a 1997 basis). The administrative payroll MOS and monthly sales MOS were also constructed using the current methodology, with appropriate strata assigned for each new employer birth that reported to the 2001 annuals. For each sampling unit considered, the stratum based on reported annual data was compared to the administrative payroll stratum and monthly sales stratum. The difference in Table 4 was calculated as the stratum determined using the reported annual sales MOS minus the stratum determined using the administrative payroll MOS. The difference in Table 5 was calculated as the stratum determined using the reported annual sales MOS minus the stratum determined using the monthly sales MOS.

Table 4: Percent Distributions of New Employer Births by Difference between Stratum Determined Using Reported Annual Sales and Stratum Determined Using Administrative Payroll

	Stratum Differences												
Sector	Equal	1	-1	2	-2	3	-3	4	-4	5	-5	6 to 12	-6 to -12
Merchant Wholesale	35.33	9.67	23.67	2.33	9.67	1.33	6.00	1.67	4.33	0.67	1.67	1.33	2.33
Retail	37.42	9.05	23.12	2.03	11.39	0.88	5.37	0.88	2.47	0.54	2.96	0.49	3.4
Service	30.81	14.45	16.65	7.63	8.03	4.28	4.57	2.02	2.60	1.33	1.73	3.64	2.26
Total	34.29	11.54	20.26	5.21	9.99	2.85	4.64	1.45	2.67	0.80	1.51	1.98	2.81

Table 5: Percent Distributions of New Employer Births by Difference between Stratum Determined Using Reported Annual Sales and Stratum Determined Using Monthly Sales

	Stratum Differences												
Sector	Equal	1	-1	2	-2	3	-3	4	-4	5	-5	6 to 12	-6 to -12
Merchant Wholesale	40.67	5.00	25.00	1.00	11.33	1.00	5.00	1.33	5.33	0.00	1.33	0.33	2.68
Retail	37.92	7.29	24.87	2.03	11.89	0.88	5.37	0.49	2.96	0.11	2.08	0.11	4.00
Service	35.84	9.56	23.07	4.17	7.86	2.14	3.76	1.27	2.95	0.58	1.62	1.85	5.33
Total	37.20	8.15	24.07	2.91	10.04	1.45	4.62	0.91	3.14	0.31	1.81	0.91	4.48

Tables 4 and 5 show little difference between the strata determined using either method or the strata based on reported annual sales. For 66.09 % of all new employer births that reported annual sales in a 2001 annual survey and had payroll data available, the stratum determined using administrative payroll and the one determined using reported annual sales were within one. For 69.42 % of all new employer births that reported to a 2001 annual survey and had SQ-CLASS data available, the stratum determined using SQ-CLASS form data and the one determined using reported annual sales were within one.

5. Conclusions and Future Research

From Section 3, there appeared to be negligible differences between the estimates, except for NAICS 44XXXX, and between the estimated standard errors, except for NAICS 44XXXX and NAICS 62XXXX, when a new sample based on only administrative payroll to determine MOS was selected for all industries. Outliers in these two industries will be looked at in greater detail to determine if these affect the estimated standard error calculations. Also, the effects of adding new employer births to only one year of the current sample were looked at in this paper. The effects of future years would have to be looked at to determine the cumulative effect these new employer births have on the estimates. A simulation study in which the second phase is replicated should also be conducted to determine which MOS methodology to use to select new employer births.

Based on the findings in Section 4, the monthly sales MOS and the payroll MOS resulted in strata with slight differences, when compared to the strata that resulted from reported 2001 annual sales. The findings in this paper suggest that the MOS determined using only payroll data produced estimates comparable to the SQ-CLASS form data MOS. Thus, if future research determines that industry classification is accurate, a one-phase design may be feasible. Implementing a one-phase design would allow new employer births to be added to the surveys approximately three to six months after a new employer birth is identified, providing a more complete coverage of the universe.

The SQ-CLASS form response is currently used to impute for nonresponse to the Monthly Retail Trade Survey (MRTS) and the Monthly Wholesale Trade Survey (MWTS). Research to compute a monthly estimate from the administrative payroll data will be conducted if only administrative payroll is used to compute a MOS. Also, it has been thought that the new employer births have little effect on the estimated month-to-month trends from MRTS and MWTS. Future research will look at the effect that new employer births have on the monthly estimates.

References

Burton, J., J. Hunt (1999), "BSR-2K First and Second Stage Sampling Specifications," unpublished memorandum, Washington DC: U.S. Census Bureau, Service Sector Statistics Division.

Burton, J., C. King, and J. Hunt (2001), "Benefits and Limitations of Using Only Administrative Data to Update Current Business Surveys with New Employer Births," *Proceedings of the Survey Research Methods Section,* Atlanta GA: American Statistical Association.

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," Proceedings on the Second International Conference on Establishment Surveys, Buffalo, NY: American Statistical Association.

Konschnik, C., E. Walker, et al. (1999), "2002 Redesign Report of Administrative Records Team," unpublished memorandum, Washington DC: U.S. Census Bureau, Economic Processing and Coordination Division.

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

Walker, E. (1997), "The Census Bureau's Business Register: Basic Features and Quality Issues," *Proceedings of the Survey Research Methods Section,* Alexandria VA: American Statistical Association.