REASSESSMENT OF THE USE OF AN AREA SAMPLE FOR THE MONTHLY RETAIL TRADE SURVEY

Carl A. Konschnik, Carol S. King and Scot A. Dahl, U.S. Bureau of the Census
Carl A. Konschnik, U.S. Bureau of the Census, Business Division, Washington, D.C. 20233

KEY WORDS: Supplementary Coverage, Business Births, Nonemployers, Survey Frames

I. Introduction

An area sample is used as a supplementary sample in the Census Bureau's monthly and annual retail trade surveys and the Service Annual Survey. The main list sample for these surveys is selected from the Census Bureau's Standard Statistical Establishment List (SSEL) which includes contributions from the Censuses of Retail Trade and Selected Services and from the administrative records of the Federal Government. The area sample is used to cover businesses not represented by the list sample. These businesses are of two types: new employer businesses (birth employers) and nonemployer businesses.

The question of whether there are viable alternatives to the area sample to cover new employer births and nonemployers for the monthly and annual retail trade surveys is a perennial one for the Business Division of the Bureau of the Census. This question is driven by two major considerations: (1) the area sample is costly to operate (about 1 million dollars per year); and (2) the area sample estimates have high variance due to the small sample size (retail establishments within the monthly area sample segments represent approximately 1/1000 of the universe of retail establishments).

The cost of the area sample can be spread to the services area also where it is used for the Service Annual Survey. Therefore, any potentially less costly alternatives to an area sample for retail must consider the cost savings (or increase) to the Service Annual Survey and other services programs as well in arriving at a total savings differential. The overall question then becomes one of whether we can produce monthly and annual estimates for retail birth employers and nonemployers which have the same (or smaller) variance as the area sample, and do the same thing for the services programs, at less than the area sample cost. Although the services survey needs constitute a very important part of the overall area sample decision, chiefly only retail will be discussed in what follows.

This paper focuses on the question of whether any alternative to an area sample can provide good (with low variance and little or no bias) estimates for the birth employer and nonemployer establishments in the Monthly Retail Trade Survey. Our preliminary conclusion is that such an alternative exists and that, once implemented, it would cost significantly less than the area sample on an annual basis. This alternative places increased reliance on government administrative records and would have a major impact on reducing processing costs and improving data quality for all monthly and annual business surveys. It could impact the economic censuses as well.

II. General Description of the Present Area Sample

We first present a general description of the area sample in relation to the list (or mail) sample. For the details of the selection of the main list sample drawn from the SSEL, see the paper by Detlefsen et al. (1991). The initial list sample is supplemented quarterly with new births which come into the SSEL. Even though the SSEL is continuously augmented for birth employers, there is a lag of from 12 to 18 months in representing birth employer businesses by the list sample. The SSEL does not contain nonemployer businesses.

For the business surveys, by birth employer businesses here we include not only businesses which are begun without any predecessors, but also those which are purchases of ongoing businesses. The primary identifier of the business firms on the SSEL is either a company and establishment number or a Federal Employer Identification (EI) number. The area sample uses an identifier related to the geographically defined Primary Sampling Unit (PSU), its segment and a line number (reflecting its position within the segment).

In recent years the area sample has accounted for a range of about 5% to 8% of the monthly retail sales estimate for the United States as produced by the Monthly Retail Trade Survey. The average annual percent of area sample contribution to total retail sales is about 5 to 6 percent. The estimated sales for nonemployers are about 3% of the total sales estimate, varying between 2% and 4%. The birth employers in the area sample account for about 3% to 5% of the total monthly sales estimate depending on the length of the delay in representing the births in the list universe. The longer the delay, the larger the percentage of sales represented by the

births in the area sample. For some retail kinds of business the area sample comprises 15 to 30 percent of the monthly retail sales estimate for that kind of business. The present area sample consists of selected segments in 12 certainty and 47 noncertainty PSUs. A historical review of the selection of the present area sample is given in Isaki, et al. (1981).

Because of the age (some noncertainty PSU segments date back to the fifties) of the current noncertainty portion of the area sample and to relieve field problems in canvassing widely different segments, the reselection of noncertainty PSUs is being considered. This, however, is a time consuming and costly effort with the field spotting costs and other processing costs estimated at about $600,000 over a 5 year period. Any new area sample would be constrained (by estimated field and processing costs) to have the same size and hence to have the same precision as the existing sample. Therefore, this is a key juncture at which to take another look at whether other alternatives are superior. The answer to this question will dictate our activity in this area for a number of years. As a partial update of the area sample, new segments were selected from the 12 certainty PSUs between 1981 and 1987. Roughly 24 percent of the retail and 30 percent of the services established fall within these PSUs.

The procedures used to tabulate the area sample data are relatively involved since any establishment belonging to a business given a chance of selection in the list sample and still active (to receive a mailing of Form 941 - Employer's Quarterly Federal Tax Return) on the Business Master File (BMF) mailing list is not to be tabulated in the area sample. Further, if these establishments can be identified ahead of time, the area sample enumerator will not collect data for them. The process of deciding whether an individual establishment should be tabulated or not involves matching the area sample establishment's EI to the list or universe file from which the list sample was selected, searching various files of employers, and finally utilizing a special decision procedure when there is conflicting information on the establishment's coverage by the list sample. This process uses a combination of both computer and clerical work. It should also be pointed out that the procedures for enumeration of the area sample segments are for a canvass of certain business establishments. Households, without any sign of business activity, are not canvassed. An additional description of the relationship between the list and area samples is given in Konschnik et al. (1985).

III. The Effects of the Area Sample on the Monthly Retail Trade Survey

The average monthly area sample contribution for the period 1982 to 1990 is 5.9 percent, with 3.1 percent due to birth employers and 2.8 percent due to nonemployers.

At kind of business levels, the area sample percent contributions vary widely. Table 1 which follows shows for the year 1990 the percent of sales contributions for several publishable kinds of business broken down by employers and nonemployers. These levels of business reflect the relatively small effect on general merchandise stores, fairly typical effects on grocery stores, and gasoline service stations, and the substantial effects on the drinking places and liquor store estimates. Table 1 is based on composite estimates which reflect the unbiased estimates from several past months in addition to the given month that the estimate represents.

The area sample effects on the trend of our estimates for 1990 are typical of other years. Table 2. highlights total retail and some kinds of business which are representative of the variation in effect. It also displays the month to month change (preliminary composite estimate for the current month to the final composite estimate for the prior month) in the estimates (unadjusted for seasonality) with (w) and without (w/o) the area sample for these businesses.

Table 3 shows the difference in the month-to-month trend in terms of standard error (of the list sample plus area sample) for the same levels as given previously in Table 2. Tables 2 and 3 show that the area sample effects on the month-to-month trends are usually small. Work is continuing on measuring and assessing the impact of the area sample on the monthly retail sales estimates. This work will provide a framework for determining the comparative benefits of other alternatives to the area sample.

IV. Alternatives to the Area Sample

IV.A. Alternatives Considered in the Past

Over the years several alternatives to the present area sample for covering employer births and nonemployers have been considered. Some of these are: 1) the use of commercial lists of businesses; 2) the use of a sample of nonemployers drawn every five years from the administrative lists used in the censuses; 3) the use of telephone canvassing of area segments instead of personal enumeration; 4) the use

Table 1. Average Monthly Area Sample Contribution to Final Retail Sales Composite Estimate for Selected Kinds of Business in 1990 (in %)

SIC	Kind of Business	Employer Sales	Non Employer Sales	Total
Total	Total Retail	2.68	2.73	5.41
53	General Merchandise Stores	0.18	0.23	0.41
541	Grocery Stores	2.20	2.31	4.51
554	Gasoline Service Stations	3.97	4.15	8.11
5813	Drinking Places	11.19	16.13	27.32
5921	Liquor Stores	4.78	11.28	16.06

Table 2. Month to Month Change in Retail Sales for Selected Kinds of Business (in %)

1990	Total		53		541		554		5813		5921	
	w/ AS	w/o AS	w/ AS	w/o AS	w/ AS	w/o AS	w/ AS	w/o AS	w/ AS	w/o AS	w/ AS	w/o AS
J	-24.8	-25.5	-62.7	-62.7	-10.9	-11.1	-3.9	-3.3	-8.3	-10.1	-36.0	-37.9
F	-4.2	-4.4	3.7	3.8	-5.0	-6.7	-7.2	-7.6	-2.6	-1.9	-2.7	-2.5
M	16.6	17.1	28.7	28.8	13.4	13.6	11.3	11.5	11.8	14.4	13.5	14.8
A	-2.5	-2.7	-1.9	-1.9	-4.6	-5.0	0.1	-0.2	-3.6	-8.1	-2.5	-2.9
M	5.8	5.8	5.5	5.5	5.6	5.7	5.3	5.5	2.7	1.6	7.8	8.5
J	-0.3	-0.4	-1.4	-1.4	0.9	0.9	2.9	2.8	1.6	0.7	3.4	3.5
J	-2.8	-2.9	-8.7	-8.7	-0.2	-0.6	1.0	1.1	-1.2	-1.5	-1.1	-2.4
A	5.0	5.1	14.5	14.6	1.9	1.6	8.6	8.7	3.2	3.5	1.9	2.7
S	-6.8	-7.0	-9.8	-9.9	-4.2	-4.3	-2.4	-2.3	-0.5	1.6	-7.7	-8.9
O	3.4	3.5	6.8	6.8	0.7	0.8	5.7	6.1	5.4	4.9	0.9	1.6
N	2.6	2.7	27.4	27.5	1.1	1.3	-2.6	-2.6	-1.1	-1.6	5.4	6.0
D	16.1	16.4	49.4	49.4	6.9	6.8	-3.0	-3.8	2.3	1.3	41.0	44.2

Table 3. Differences in 1990 Unadjusted Month to Month Changes With and Without the Area Sample (in Standard Errors)

SIC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Total	3.4	0.6	-1.4	0.7	-0.1	0.3	0.2	-0.1	0.7	-0.4	-0.3	-0.7
53	1.0	-0.7	-0.9	0.0	-0.6	-0.0	-0.2	-0.7	1.2	-0.2	-1.2	0.1
541	1.0	6.0	-0.8	1.4	-0.5	0.1	1.1	0.8	0.2	-0.4	-0.6	0.3
554	-2.0	1.2	-0.5	1.1	-0.5	.1	-0.4	-0.2	-0.4	-1.4	0.1	2.7
5813	2.4	-0.9	-2.9	5.9	1.4	1.1	0.3	-0.3	-2.7	0.5	0.5	1.2
5921	4.9	-0.2	-1.9	0.7	-1.1	-0.2	2.2	-1.4	2.1	-1.3	-0.9	-3.7

of the Current Population Survey (CPS); and 5) the use of factors developed from previous periods. These methods along with the drawbacks of each are discussed in Isaki et al. (1981).

An earlier review of alternatives to an area sample by Turbitt, et al. (1972) discussed using a complete list sample approach or the use of factors. While their investigation into using factors showed that approach to produce large errors, they concluded that a list sample design exclusively could possibly be used to represent employer births and nonemployers. Their concern was largely one of the cost (relative to the area sample cost) of manipulating the huge income tax files, but they also noted the problem of any remaining lag in representing very recent employer and nonemployer births.

Recently, administrative record processing efficiencies as well as data processing improvements in general have provided some alternatives along the lines envisioned in the preceding paragraph. These make it possible to significantly reduce the lag time for introducing birth employers into the surveys and to construct a sample of nonemployers for monthly survey use and keep it updated on an annual basis.

IV.B. A Proposed Alternative to Account for Birth Employers

Sampling for employer births is presently done quarterly using a two phase operation. In the first phase a sample of approximately 8% of the new employer births is selected from the SSEL based on payroll or expected employment (or both) and an SIC assigned by the Social Security Administration (SSA). Unclassified cases are sampled as a separate group. Selected EIs are mailed a report form which asks for two months sales, company affiliation, new or refined SIC information, etc. Based on the results of this birth survey mailing, a more accurate SIC and a sales measure of size are used in the second phase of sampling done three months after the first phase. The selected births (about 20% of the first phase selects) in this second phase are panelized and represented initially in the list sample for the data month in which they are sampled in the second phase.

Adding to the three months lag between the first and second phases of the double sampling procedure is another 9 to 15 months on average between the time cases start operating under their new EI number and our first phase of sampling. This lag is caused by several factors: the time between when businesses start operating and their filing an IRS Form SS-4 (application for an EI number); the time it takes for a copy of the SS-4's to be sent from IRS to SSA for SIC coding; and the time it takes for the SSA to assign SIC codes and give them to the Census Bureau. Another source of lag is caused by many EI numbers showing up as reporting nonzero payroll to IRS for which no SIC information is obtained by the Census Bureau. More than a full year (or four quarters of payroll) is spent waiting for an SIC code from SSA before these new EIs are mailed in the first phase as unclassified. Roughly half the employer births fall into this "payroll only births" category, which contributes significantly to the 12 to 18 months overall lag.

Clearly, without an area sample, the lag in representing employer births in the list samples would have to be as small as possible--ideally none, but this probably could not be achieved. Some steps we might take to reduce the current lag in representing births are both 1) and 2) which follow below.

1) Select a sample of EI numbers from the population of EIs not yet subjected to sampling as soon as they a) have nonzero payroll reported for the most recent quarter or b) become BMF active (have an IRS Form 941, "Employer's Quarterly Federal Tax Return", active filing requirement). This means that the IRS is mailing a Form 941 to the EI with an expectation that the EI has employees and will need to file the quarterly report.

This procedure would result in most EI births being selected as "unclassified" as to SIC and would substantially increase the number of cases to be mailed a birth form.

2) Select births on a monthly rather than quarterly basis. This would require a first phase and a second phase of sampling as before. Turnaround of the mailing and followup would have to be done on monthly rather than quarterly cycle. A monthly cycle was used prior to 1977. It would be possible to tabulate the returned birth survey forms as part of the monthly estimates in the month they are returned to further reduce the lag.

The impact of these changes would be: 1) the reduction of the lag in representing birth employers; and 2) the increased cost of mailing, following up and processing the birth survey, due in part to the fact that some potential employer births may be mailed in the survey but are found to be not operating.

In addition to reducing the lag for representing employer births to a minimum,

some steps would be necessary to account for those new births that can't be covered. Just how small the contribution of uncovered births is and how well this imputation procedure accounts for them will have to be measured.

IV.C. A Proposed Alternative to Account for Nonemployers

Nonemployers are represented in the retail census through tax return records received from the IRS. These records include the Form 1040, Schedule C returns for sole proprietorships along with Forms 1120 or 1120S returns for nonemployer corporations and Form 1065 returns for nonemployer partnerships. For the 1040-C returns, the Social Security Number (SSN) is the primary identifier and the EI number is asked for those that have one. For partnerships and corporations, the EI is the primary identifier. Nonemployer corporations and partnerships are identified as those EIs having nonzero receipts but having zero payroll on the SSEL. Nonemployer sole proprietorships are identified as having nonzero 1040-C receipts but zero payroll on the SSEL for their associated EI number. Some sole proprietor employers do not provide their EI number on the 1040-C as requested. In this case certain procedures are used to distinguish 1040-Cs which are employers from those which are not. These procedures, some of which were used for the 1987 censuses of retail and services, are described in Konschnik and Moore (1990).

The same or similar methods as those used to identify retail nonemployers for a census year can be used in any year. Thus, we could construct a list of nonemployers and update it annually. This list could be used to draw a sample of nonemployers which could be canvassed by mail for monthly data. Imputation methods similar to those for employers could be used to account for new nonemployer births since the list was last updated. Methods to evaluate any resulting biases due to these new nonemployer births need to be set up for the monthly retail survey. Factors developed from a prior annual period could also be evaluated as an alternative to a sample for representing nonemployers.

IV.D. Phase-in of the New Procedures

The recommended approach for considering the replacement of the area sample with the procedures outlined in the two previous sections is to phase in the new procedures while continuing to conduct the monthly area sample. This strategy will allow direct comparisons of both methodologies relative to variance and bias. The area sample could also measure the bias due to the lag in representing employer and nonemployer births in the list samples. We would be able to drop the area sample only after we are convinced that the new method will give improved estimates at less cost and that the risks of not having the insurance provided by the area sample against processing errors or other list deficiencies are minimal.

Replacement of the area sample would best be preceded by and include the following phases.

1) Change the birth sampling procedures to identify and sample employer births as soon as possible, selecting them as unclassified with respect to kind of business if necessary. Process births on a monthly rather than quarterly cycle. This would have the effect of representing more cases on the list sample basis rather than the area sample basis and should begin to lower the overall variances and reduce the effects of area sample panel differences.

2) Begin to receive the IRS Form 1040-C, 1065, 1120 and 1120S on an annual basis and process receipts data onto the SSEL. Identify the universe of nonemployers and evaluate the quality and completeness of industrial classification.

3) Use the administrative records data obtained in 2) to represent nonemployers for the annual surveys of retail and services, perhaps augmented by a supplemental survey of unclassified nonemployers drawn from the list. This would put the annual survey processing of nonemployers on a basis similar to the censuses and would especially benefit the services estimates of nonemployers since many services businesses are "nonvisible" to our current area sample enumerators.

4) Prepare an additional alternate sample design for BSR-97 in early to mid 1994 which assumes no area sample usage. This design would produce tentative sample sizes for employers as well as nonemployers.

5) Begin an overlap processing using the new procedures while continuing to conduct the area sample. If the new procedures prove successful, the area sample could be dropped perhaps as early as the latter part of 1997. During this time various alternative adjustment procedures to account for nonemployers and lags in birth employers could be evaluated.

IV.E. Issues to Be Addressed with the New Procedures

Many issues with the new list sample approach must be addressed. One of these is the process of achieving complete, unduplicated coverage of the employer and nonemployer universes each month. Currently, a selected noncertainty EI which becomes BMF inactive is dropped from the list sample. However, an EI can become BMF inactive not only when it goes out of business, but also when it continues to operate but no longer has any employees. Inactive EIs which are still operating a business and are dropped from the list sample are covered by the area sample. In the new procedures, such EIs, once selected into the list sample would remain represented by the list sample. In effect, BMF inactive EIs would continue to be canvassed unless and until they stopped operating a business. In this sense then, the employer list sample would represent nonemployers as well and the nonemployer list sample would initially be drawn from a supplementary list of all businesses not covered by the employer list. Since the SSN is the primary identifier for sole-proprietorship nonemployers, this number would have to be requested for each sole-proprietorship birth EI so that we don't sample as an employer birth a business previously represented in the nonemployer SSN universe. In this sense then, the nonemployer universe could contain employers. Effective unduplication rules will have to be carefully worked out.

A key component of any unduplication will be the ability to match sole-proprietorship SSNs with their associated EI number. IRS has taken steps to have the sole-proprietorship's SSN included on the BMF file. These steps along with better reporting of any associated EI numbers on the 1040-C would greatly aid our unduplication work. Clearly, our survey questionnaires would need to request the SSN from any selected sole-proprietorship EI sampling unit, and the EI number would be requested of any selected SSN sampling unit to also aid in unduplication.

Other possible problem areas which must be considered in a decision to adopt the new procedures include: 1) the quality of the kind of business coding on the tax returns—a study is being done to assess the quality of the kind of business coding on the 1040-C returns from 1987 (codes assigned by taxpayers are being compared to codes assigned through a survey mailing); and, 2) the completeness, accuracy, reliability and timeliness of the administrative records. Without an area sample the retail estimates become vulnerable to any problems of completeness or accuracy in the IRS administrative records. The reliability of the administrative records source and the timely receipt of the necessary files also would be crucial. Events such as that which occurred during 1978 when IRS suspended giving 941 payroll and BMF files to the Census Bureau for a period of about 9 months would cause very serious problems with the retail estimates.

References

Detlefsen, Ruth E., and Veum, Carol, (1991), "Design Issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census", presented at the 1991 annual meeting of the American Statistical Association.

Isaki, C., Konschnik, C., and Monsour, N., (1981), "Reselection of an Area Sample for the Retail and Service Surveys", Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 116-121.

Konschnik, C., Monsour, N., and Detlefsen, R., (1985), "Constructing and Maintaining Frames and Samples for Business Surveys", Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 113-122.

Konschnik, Carl A., and Moore, Richard A., "EC-14, A Study of the Methodology for Removing Employer Duplicates from the Nonemployer Universe for the 1987 Censuses of Retail and Services", (September 19, 1990). Unpublished memorandum.

Turbitt, J., Shor, M., and Woodruff, R., "20 Percent Area Sample Survey - Business Division", (September 11, 1972). Unpublished memorandum.