CHANGES IN SAMPLING UNITS IN SURVEYS OF BUSINESSES

Jock Black
US Census Bureau

1. Abstract

This paper discusses methods to handle changes in establishment-based sampling units that can have a large and possibly erroneous influence on estimates. Changes can be in classification, size, constituency of the sampling unit, or business operation. Methods used to correct the problem depend on when the change occurs and when knowledge of the additional information is gained, the magnitude of the change, the complexity of the solution and other factors. The problem is discussed in the context of the Monthly Retail Trade Survey.

2. Introduction

The Service Sector Statistics Division of the Bureau of the Census conducts periodic sample surveys of businesses to estimate important characteristics of the economy of the United States. The sampling units employed in these surveys are either individual establishments or clusters of establishments under common ownership. Commonly, a stratified sample design is used with strata defined both by size of establishment and kind of business (KB). Sampling efficiency coupled with the skewed nature of the universe of businesses results in a wide range of sampling fractions. As with any sample survey, the accuracy of the survey estimates relies on the ability of the selected sample to represent the population. For these periodic surveys, the accuracy of the sample estimates depends on the correctness of the kind of business classification; the measure of size both at the time of the initial sample design and as the survey is carried forward through time; and the accurate identification of ownership relations. Because the samples used for these surveys are reselected at five-year intervals and the business universe is constantly changing, the initial information used for sample design and estimation may occasionally be revised. When this happens, corrective procedures may be undertaken in the hope that the sample estimates will more accurately reflect the current population's parameters.

This paper will discuss various changes that can occur to sampling units during the survey process and several methods used to address the changes. Different methods may be adopted depending on the nature of the change, the consequences of different decisions, and whether or not the unit is self-representing. Although the methods may be appropriate for many surveys, this paper will discuss the problems in the context of the Monthly Retail Trade Survey (MRTS).

3. Monthly Retail Trade Survey Design

The United States Bureau of the Census conducts the Monthly Retail Trade Survey (MRTS) to estimate retail sales in total and by specific kinds of business (KBs) for employer and nonemployer businesses operating in the United States. Each month, estimates of monthly levels and changes in levels from a month ago and a year ago are published. Retail sales are a primary measure of household demand and changes in them are widely followed as the most timely indicator of broad consumer spending patterns.

The sample design of the MRTS is similar to the design of the sample used for the Annual Retail Trade Survey (ARTS). A description of which is documented in the Annual Benchmark Report for Retail Trade. Further details can be found in papers by Kinyon, et al. A brief summary of the design features that are important for this discussion is given next.

1. Sales for retail establishments under common ownership are aggregated to a company level. The company unit is assigned the kind of business (KB) that accounts for the largest portion of sales.
2. Sales for retail establishments that use the same Employer Identification Number[1] (EIN) are aggregated to the EIN level. The EIN unit is assigned the KB that accounts for the largest portion of sales.
3. For each KB, the largest company sampling units are identified as certainty units (or self-representing) and EIN units associated with them are removed from the units defined in step 2.
4. The remaining universe of EIN units is stratified by KB and approximate annual sales. A second measure of annual sales, based on payroll, is used to determine total sample size and its allocation to the measure of size strata. A simple random sample without replacement (SRSWOR) is selected in each stratum.
5. Data for nonresponding units are imputed based on responses from units of the same size and KB.
6. Level estimates are computed as the weighted sum of reported and imputed data. A Horvitz-Thompson estimator is used.
7. The initial sample is updated quarterly. Lists of new EINs are received via administrative sources and are subjected to a double sampling scheme. A large first phase sample is enumerated by mail, obtaining information necessary to assign the unit to the initial sampling strata. Using this information, the second phase sample is selected systematically at the same sampling rate as the initial sample.
8. Monthly KB level estimates are benchmarked to the Annual Retail Trade Survey (ARTS) and the Retail Census. The benchmarking procedure revises the monthly level estimates so that the sum of the monthly estimates for a given calendar year is equal to the annual total derived from either the ARTS or the Census. There is about an eighteen-month period between the end of the reference year and the publication of the annual estimates. Consequently, at least eighteen monthly estimates are not benchmarked to the appropriate annual total. These estimates are revised by multiplying them by the ratio of the revised-to-sample estimate for December of the last benchmark year. For example, in May 2001, the sales estimates from the 1999 ARTS were the last year available for benchmarking.

4. Addressing Changes Involving a Single Sampling Unit

This paper will address the following changes that can occur to business sampling units: KB classification; size; mergers and acquisitions; splits (or divestitures); deaths; and births. In this section we address changes that affect an individual sampling unit and do not increase or

[1] An Employer Identification Number (EIN) is also known as a federal tax identification number, and is used by the Federal Government to identify a business entity for purposes of payroll tax withholding. It is legally required for an employer business to have an EIN.

decrease the number of sampling units in the universe – changes in classification or size of a particular unit.

In the discussion that follows it is important to note that the approach taken to address the problem is sometimes a compromise between unbiasedness, mean square error, magnitude of the problem, and ease of implementation of the solution. It is also important to note that when biased procedures are followed it is because an industry or survey expert has concluded that following an unbiased procedure will lead to grossly inaccurate measures.

4.1 Changes in Kind of Business Classification

This occurs when a sampling unit is engaging in a different principle activity than when it was subjected to sampling.

Consider the following table:

Table 1.

		Tabulated Kind of Business				**True Total**
		1	**2**	**...**	**L**	
True Kind of Business	**1**	$p_{11}\ Y_{.1}$	$p_{12}\ Y_{.2}$	...	$p_{1L}\ Y_{.L}$	$Y_{1.}$
	2	$p_{21}\ Y_{.1}$	$p_{22}\ Y_{.2}$	...	$p_{2L}\ Y_{.L}$	$Y_{2.}$
	¼	...	...	...	...	...
	L	$p_{L1}\ Y_{.1}$	$p_{L2}\ Y_{.2}$	...	$p_{LL}\ Y_{.L}$	$Y_{L.}$
Tabulated Total		$Y_{.1}$	$Y_{.2}$	...	$Y_{.L}$	

Define Y_{ij} = the sales of units whose true KB is i but which are tabulated in KB j and p_{ij} = the proportion of sales of units tabulated in KB j that should be moved to KB i. Note that only the last row of the table is observed. We desire that the estimated total sales level for a KB equal the true total sales, i.e., $\sum_j p_{ij} Y_{.j} = \sum_i p_{ij} Y_{.j}$.

Larger units do not often change their primary activity. Because of this and their importance, they are usually classified correctly. However, when their KB classification is inaccurate, equality of the two totals is unlikely because such units are essentially outliers. Thus, we have adopted the rule that certainty units should be reassigned to their proper kind-of-business. The timing of the reassignment may be postponed until benchmarking so that historical estimates are also corrected and to avoid creating a level shift in each of the KBs affected by the reassignment.

The classification of smaller units, while still important, is not as thoroughly reviewed as it is for larger units. It is reasonable to assume that for most KBs the chance of being misclassified is approximately equal, that is $p_{ij} = p_{ji}$. Restricting the table above to size strata it is reasonable to conclude that the off-diagonal entries are approximately constant and the tabulated totals are close to the true totals. This is only one reason why once a noncertainty unit enters the sample, we seldom change its tabulated KB. Other reasons are additional processing to make the changes increases the likelihood of errors, moving units has the potential to increase the variance of estimates and to create sudden, false shifts in levels and month-to-month change estimates.

4.2 Changes in size

A common occurrence for surveys of businesses that maintain sampling units over time is the growth or shrinkage of a noncertainty unit's sales during the survey period. While it is tempting to treat such units as edit failures for which a less troublesome imputed value will be created, or to Windsorize their reported data, we do not. The data for these units have been verified to be correct and genuinely larger (or smaller), are expected to maintain their growth, and come from the same distribution as the remainder of the sample. Thus, while the Horvitz-Thompson estimates are unbiased, such units will (appropriately) increase the estimates of variance and contribute to statistically significant estimates of change when, in the opinion of subject matter experts, they should not.

As an example, suppose an EIN was assumed to have about $500,000 in annual sales, or about $41,600 per month. With this level of sales, the unit was selected with a weight of 440. In its first five monthly reports, it always reported more than $130,000, once reporting over $600,000 in sales. Over the 5-month period, the reported sales accounted for about $1.1 million. If its sales continued at the reported level, the unit's annual sales would be about $2.8 million – five times the size used for initial sampling.

We have considered a number of methods to address such units including augmenting the sample with additional units, alternative estimators, and assigning new weights to all selected units in the same stratum. However, because of the short time to resolve problems and the complexity of other solutions, each of which makes some assumptions about the uniqueness of the problem, we have frequently chosen to adjust the weight on only the problematic unit.

4.2.1 Procedure to Adjust Sampling Weight

The method attempts to set the contribution of the problematic unit equal to the unit's assumed true contribution in the universe. It is convenient to interpret the weight of 440 to mean the unit represents itself and 439 other units of roughly the same size and classified in the same KB that were not selected in the sample. Let

w_{orig}	=	the weight with which the unit was originally selected,
w_{new}	=	the weight to be determined,
m_{orig}	=	the original measure of size,
m_{actual}	=	the measure of size derived from reported monthly data, and
ρ	=	the assumed proportion of units in the universe that the problematic unit is representing that have behaved like to the observed unit.

An example will illustrate how to set and interpret the rho (ρ) parameter. The unit under consideration has a weight of 440. It represents 439 other sampling units in the same size and KB stratum. If only 44 of those sampling units have grown like the problem case, we would set $\rho = 44/439 = 0.10$. Note that we do not know what the true value of ρ is because we never observe the numerator. However, if the sample in the problem unit's stratum is large enough, it is possible to estimate the proportion of units that have changed using the selected sample. To be conservative, we often set it to 1/3 as a compromise between the desired proportion of 1 and the unbiased proportion of $1/w_{orig}$.

Our constraint can be stated as

Recalibrated New Contribution	=	Assumed True Contribution in Universe		
	=	Contribution of Units That Grew	+ Contribution of Units that are Still the Original Size	

Symbolically,

$$w_{new}\, m_{actual} = \rho w_{orig}\, m_{actual} + (1 - \rho)\, w_{orig}\, m_{orig}$$

This gives $w_{new} = w_{orig}\, (\rho + (1 - \rho)\, m_{orig} / m_{actual})$.

4.2.2 Criterion for Changing Weight

Making changes that reflect new information can require a significant amount of processing work and create unnecessary confusion and revisions for data users. If the changes were to have only a negligible effect on published estimates, it is preferable either to not make the changes or postpone them until the next benchmarking. Therefore, one criterion we use to determine whether estimates should be revised is the following. If the absolute value of the effect of changing an estimate exceeds a pre-selected multiple of the standard deviation of the estimate, then the estimate should be changed. Stated differently, if the net effect of the change does not exceed the range of sampling error, then the change should not be made.

5. Changes in the Number of Sampling Units

This section discusses events that change multiple units in the universe of sampling units – business births, deaths, mergers and splits. It should be pointed out that not all births and deaths result in changes in the number of sampling units. Because the sampling units in the MRTS are clusters of establishments, births or deaths of establishments associated with existing sampling units will be accurately represented via the original sampling unit. For example, since a company sampling unit is requested to report the sales for all of its establishments, the opening or closing of individual establishments will be captured via the company's report. Thus, births and deaths of establishments are only problematic when they involve entirely new EINs or companies.

5.1 Births

The treatment of births is described briefly in item 7 of section 3 above. More information on new employer births in the sample can be found in the description of survey methodology contained in the Annual Benchmark Report for Retail Trade.

5.2 Deaths

Deaths occur when a company or EIN is no longer used for business purposes and does not have an identified successor. The treatment of certainty units differs from the treatment of noncertainty units. Company and certainty EIN deaths are rare and generally take place over a

long period so that the estimates will decrease gradually. Because these units are self-representing, no action is necessary. That is, their sales will eventually be tabulated as zero when they cease reporting.

Noncertainty EIN deaths are common and the procedures to account for them are uncomplicated. Because of the lag that occurs between the "real world" events of business births and business deaths and their official recognition on administrative data files, not accounting for deaths would introduce a downward bias in the estimates. To see this, consider a simple example. Suppose the universe consists of one EIN that goes out of business at the same time a new EIN with the same level of sales starts operation. The EIN that is out of business will stop reporting immediately and the estimate of sales will become zero until the birth EIN is identified and represented in the sample. The true sales, however, are unchanged. Therefore, it is necessary to account for deaths during the lag period until births are represented in the sample. In the MRTS, sales are imputed for EINs that go out of business until the unit is identified as out-of-business on administrative files. This assumes that a.) the time necessary to identify and process deaths is the same as the time necessary to identify and process births and b.) the sales attributable to new unsampled EINs essentially replace the sales of EINs that go out-of-business.

5.3 Divestitures and Splits

When a company divests itself of some subset of establishments and the subset(s) then becomes a separate legal entity, four possibilities exist with regard to EINs:

1. All old EINs are voided and new EINs are obtained for the successors.
2. All old EINs are retained and split amongst the successor companies.
3. One successor retains the old EINs while the other successors obtain new EINs.
4. One or more of the successor companies retain some old EINs and obtain some new EINS.

The question that arises is how these situations should be handled with regard to weighting and estimation. The answer depends on both the birth sampling process and the estimation process. It is important to note that whenever possibilities 1 or 3 occur, the new EINs will be eventually subjected to sampling as part of the birth procedures.

Many sampling procedures could be defined to deal with the problem of independent splits. This paper will consider three. To describe these assume the original universe consisted of N companies from which a sample, s, of fixed size, n, was selected. Suppose later a company C splits into two independent successors, S1 and S2.

Three possibilities are:
1. The successors are to be in the sample if the original company is in the sample. The successors are not in the sample if the original company is not in the sample.

2. If the original company is in the sample, then select only one of the successors with probabilities p_1 for S1 and p_2 for S2. If the original company is not in the sample, then neither successor will be in the sample.

3. Treat the original company as a death and treat S1 and S2 as births.

Regardless of the rule followed, unbiased estimators of total sales exist for each of these rules. For example, for the first option, we have

$$\hat{Y} = \sum_{j=1}^{N} \frac{y_j'}{p_j} a_j$$

where y represent sales; $\alpha_j = 1$ if unit j is in the sample, = 0 otherwise; π_j = probability that unit j is in the sample; and $y'_j = y_{S1} + y_{S2}$ for unit C, y_j otherwise. Clearly $E[\hat{Y}] = Y$ and

$$Variance(\hat{Y}) = \sum_{j=1}^{N} \left(\frac{y'_j}{p_j} \right)^2 p_j (1 - p_j) + \sum_{j \neq k}^{N} \left(\frac{y'_j}{p_j} \right) \left(\frac{y'_k}{p_k} \right) (p_{jk} - p_j p_k).$$

For the second option, define $\hat{Y} = \sum_{j=1}^{N} \frac{y''_j}{p_j} a_j + \frac{y_{S1}}{p_{S1}} a_{S1} + \frac{y_{S2}}{p_{S2}} a_{S2}$, where $\pi_{S1} = p_{1*}\pi_C$ and $\pi_{S2} = p_{2*}\pi_C$ and $y''_j = 0$ for unit C, y_j otherwise. Clearly $E[\hat{Y}] = Y$. Calculating the formula for the variance of $\hat{Y}$ is straightforward.

Essentially, this option subsamples C via its successors to estimate the original unit's sales contribution. However, since only one successor is chosen this option does not allow for a design-based estimator of variance from the sample.

For the third option, no change in the usual Horvitz-Thompson estimators is needed. However, care must be taken to ensure that the successor EINs actually acquire new EINs and are subjected to sampling.

In practice, when a noncertainty EIN splits, possibilities 1 and 3 above are most common. In this case, no special action is required. When a certainty sampling unit splits, all successor companies are identified and added to the sample as certainty companies. When a certainty sampling unit divests itself of establishments, research is undertaken to determine if the establishments were acquired by a sampling unit that has been subjected to sampling. If a large proportion of the original company was divested, all companies that acquire these establishments are made certainty units.

5.4 Mergers and Acquisitions

Mergers and acquisitions involve reductions in the number of sampling units. A merger occurs when at least two sampling units, P_1 and P_2 (the predecessors), combine to form a single sampling unit, S (the successor), with a different identifier – either a new EIN or a new company. An acquisition occurs when the successor maintains the same identifier as one of the predecessors. To analyze this situation it is convenient to break the analysis into pieces according to the status of the predecessors - selected with certainty, selected as a noncertainty, and not selected. Regardless of the situation, we consider the following estimator of total after the merger given by $\hat{Y} = \sum_{j=1}^{N} a_j w_j \tilde{y}_j + a_S w_S y_S$.

Where $\alpha_j = 1$ if unit j is in the sample, = 0 otherwise; w_j = weight that unit j would be tabulated with in the sample; and $\tilde{y} = y_j$ for units other than P_1 or P_2, and = 0 for P_1 and P_2. Note that this estimator is unbiased if $w_j = 1/p_j$, where p_j = the probability that unit j is selected.

Examining the estimated total before and after the merger will be used as a guide for decisions. Ideally, we would have

$$\left[\sum_{j=1}^{N} a_j w_j \tilde{y}_j + a_S w_S y_S \,\middle|\, P_1\ P_2\ \textit{status, S not in U}\right] = \left[\sum_{j=1}^{N} a_j w_j \tilde{y}_j + a_S w_S y_S \,\middle|\, P_1 \textit{ and } P_2 \textit{ not in U, S in U}\right].$$

Since in some instances, it might be desirable to subject S to sampling and introduce a stochastic component, we need to consider the expected value of the right hand side of this equation. Six possibilities are analyzed below.

5.4.1 P_1 certainty, P_2 certainty

In this situation, the contribution to the estimator prior to the merger from the two predecessors has two terms contributing $y_{P1} + y_{P2}$. After the merger, the estimator tabulates P_1 and P_2 with zero sales and includes a term $(w_{Sy}y_S)*E[\alpha_S]$ for the successor, where $y_S = y_{P1} + y_{P2}$. To maintain equality, we must have $w_S = 1$ and $E[\alpha_S] = 1$. This occurs when $\pi_S = 1$. In words, we include the merged company in the sample as a certainty sampling unit. The same holds in the case of an acquisition.

5.4.2 P_1 Certainty, P_2 Selected Noncertainty

In this situation, equating before and after estimates gives $y_{P1} + w_{P2}y_{P2} = w_S y_S E[a_S]$.

If we were to require $E[\alpha_S] = 1$ and $w_S = 1$ we find the new sample underestimates the sales by $y_{P2}*(w_{P2} - 1)$. One way to preserve equality is to include the combined unit in the sample with a weight, $w_S = (y_{P1} + y_{P2})/(y_{P1} + w_{P2}y_{P2})$.

Strictly speaking, a subtle change in the expectation of the estimator would occur before and after the merger or acquisition when a noncertainty unit is acquired, assuming the sampling was performed on the actual frame and weighting was based on selection probabilities. The universe count would decrease by one and the sample size would either decrease by one or stay the same. This would have the effect of changing the selection probabilities and the consequent weighting. For example, in the case of SRSWOR, the before and after estimated totals are

$$\hat{Y}_{Before} = \frac{N}{n}\sum_{j=1}^{n-1} y_j + \left[\frac{N}{n} y_{P2} + y_{P1}\right] \text{and } \hat{Y}_{After} = \frac{N-1}{n-1}\sum_{j=1}^{n-1} y_j + \left[y_{P2} + y_{P1}\right].$$

Note that the contribution to the estimate from units not involved in the merger increases, while the contribution from terms involving P_1 and P_2 decreases. Implementing this option, however, would require reweighting all noncertainty units in the same stratum as P_2 at the time of initial sampling.

Frequently, though, y_{P1} is much larger than y_{P2} and w_S defined above is close to one. As a simple procedure, we add the successor to the sample as a certainty. A rationalization of this approach, which, arguably, could lead to an underestimate is as follows. The noncertainty unit in our sample represents other EINs not in the sample that have been or will soon be acquired by certainty companies. The acquiring certainties have been reporting sales for their acquired EINs. Since these EINs are still being represented in the sample by non-acquired EINs, there has been an increasing overestimate. When the selected noncertainty unit is acquired, the estimates are decreased to a more accurate level.

An alternative method to maintain unbiasedness is to artificially split the merged unit into its component parts by using the following procedure. Based on the most current reported information, estimate the ratio of each predecessor's sales to the total sales of the successor. That is, calculate $r_1 = y_{P1}/y_S$ and $r_2 = y_{P2}/y_S$. For subsequent months, create two parts from the successor unit with sales of $r_1 * y_S$ and $r_2 * y_S$. Tabulate one part with weight $\pi_{P1} = 1$ and the other with weight π_{P2}.

5.4.3 P_1 Certainty, P_2 Not Selected

The third situation is similar to the second except that there is no contribution to the estimator from P_2. Requiring equality leads to weighting the combined unit with a weight $(y_{P1} + y_{P2})/(y_{P1})$. Similar analysis of the before and after estimates would show that the contribution to the estimate from units not involved in the merger decreases, while the contribution from terms involving P_1 and P_2 increases. In practice, our procedure ensures that the merged unit is a certainty in the ongoing sample and prevents either of the predecessors from remaining in the sample.

The alternative used in the second situation can be applied here. In this case, we set r_2 = zero and proceed as above.

5.4.4 P_1 Selected Noncertainty, P_2 Selected Noncertainty

To maintain equality of the before and after estimates in this situation, we desire the combined unit have a weight of $(y_{P1} + y_{P2})/(w_{P1}y_{P1} + w_{P2}y_{P2})$. In the case of a merger in which a new EIN is obtained for the combined unit, it would be possible to treat the two predecessors as deaths and subject the new unit to sampling with this probability. However if one of the original EIN units is used for the combined unit this is not possible. If an acquisition occurs one EIN will remain in the sample and the other will eventually be tabulated with zero sales. If the original sampling weight is maintained, we would replace the contribution of the two units, $(w_{P1}y_{P1} + w_{P2}y_{P2})$, with the weighted contribution of the combined unit $w_S(y_{P1} + y_{P2})$, introducing a slight bias in the estimates.

Without going through the details, the alternative introduced in 5.4.2. can be used here to create two reporting parts, and the sales of the combined unit can be decomposed and weighted using the original weights to maintain unbiasedness. Frequently in the MRTS, both units are given a special status to indicate that their reported data is to be overwritten with imputed values. The effect of this is to maintain both units in the sample as they were originally sampled.

5.4.5 P_1 Selected Noncertainty, P_2 Not Selected

As in 5.4.4, we can estimate the contribution to S from P_1 and apply it to the combined report. The amount $r_1 * y_S/\pi_{P1}$ is then tabulated.

5.4.6 Neither P_1 nor P_2 in Sample

Since neither unit contributes a nonzero amount to the conditional expectation, no change in sampling or estimation is required for this merger. Presumably, the noncertainty sample represents these units whether separately or together.

6. Summary

During the period in which a sample is used, many changes can occur to the units in the sample. If neglected or ignored, the changes could lead to significant inaccuracies in published estimates. Adjusting the sample or sample procedures to account for changes is seldom an easy task but necessary. In deciding how to handle changes, the survey designer must take account of both the statistical methodology and the flexibility of the processing systems and propose a solution that may not be ideal for either, but is practical for both.

The author would like to thank William C. Davie Jr. for his suggestions and comments in the preparation of this paper.

7. Bibliography

Cochran, W. (1977), *Sampling Techniques*, New York: John Wiley & Sons.

Isaki, C., "Rules for Maintaining Frame Over Time," Unpublished Note, 1975.

King, C., "Status Change Rules for the Annual Surveys," Internal Census Bureau Memorandum, BSR-2K-4-P-1, 2001.

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.

U.S. Census Bureau (2000), Current Business Reports, Series BR/99-A, *Annual Benchmark Report for Retail Trade: January 1992 to December 1999*, Washington, DC.

Wolter, K., "Some Thoughts on Mergers and Acquisitions," Unpublished Note, 1975.

Wolter, K., "Some Thoughts on Splits and Successors," Unpublished Note, 1975.