DE GRUYTER OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 773–785, http://dx.doi.org/10.2478/JOS-2014-0047

On Precision in Estimates of Change over Time where Samples are Positively Coordinated by Permanent Random Numbers

Annika Lindblom[1]

Measures of period-to-period change are key statistics for many economy surveys. To improve the precision of these estimates of change, the majority of the business surveys at Statistics Sweden select stratified simple random samples (STSI) at different points in time, ensuring positive correlation between samples (overlapping samples) by using permanent random numbers (PRN). Statistics Sweden normally selects positively coordinated STSIs drawn from an updated Business Register (BR). In these samples, the industry strata are usually stratified further within industry into size strata. When the most recent sampling frame contains updated classification variables for all units, enterprises can change stratum between two sampling occasions. A drawback of the coordinated sample selection procedure is that the desired correlation between the two samples decreases if the proportion of enterprises that change strata is substantial. Consequently, the sample design must anticipate the potential effect of stratum changes between samples. This article presents a study that examines how the design of a repeated business survey affects the precision in estimates of change over time using the Turnover in the Service Sector survey conducted by Statistics Sweden as an example.

Key words: Measures of change; sample coordination; survey design; variance estimation.

1. Introduction

An important issue in many repeated business surveys is to determine whether the period-to-period change in an estimated total is statistically significant. To improve the precision of estimates of change, the majority of the business surveys at Statistics Sweden use samples from separate points in time ("sample occasions") that are positively coordinated (overlapping) by permanent random numbers (PRN). This positive coordination over time introduces dependence between the obtained samples, inducing positive correlation between the two level estimates, which in turn increases the precision in estimates of change over that obtained from independent samples.

Statistics Sweden normally uses positively coordinated stratified simple random samples (STSI) drawn from an updated Business Register (BR). The stratification is usually performed by industry, further grouping units within an industry into size strata. One drawback of this coordinated sampling procedure is that the desired correlation decreases between the two samples if the proportion of enterprises that change strata

[1]Senior Methodologist, Statistics Sweden, 701 89 Örebro, Sweden. Email: annika.lindblom@scb.se
Acknowledgments: I would like to thank Dr. Lennart Nordberg for helpful discussions during the course of this work which have led to substantial improvements. I would also like to thank the Editor, the Associate Editor and three referees for their useful review of earlier versions of this article.

is substantial. The sample designer must therefore anticipate the potential effect of stratum changes between samples. A detailed size stratification procedure (creating numerous small strata within a given industry) promotes high precision in each level estimate but often results in a smaller overlap (less correlation) between samples. On the other hand, a less detailed stratification (allowing wider ranges within size strata) yields less precise level estimates but a larger overlap (higher correlation) between samples.

Despite the fact that coordinated samples are commonly used at Statistics Sweden, there has been little research conducted exploring the "tradeoff" between the usage of a detailed stratification and size of overlap on the precision of estimates of change. Such knowledge would be very useful for future sample designs. This article presents the results of a simulation study conducted at Statistics Sweden that compares the precision of the same change estimates obtained by using three different STSI sampling designs selected from the frame data of the Turnover in the Service Sector survey (hereafter referred to as the TSSS). This study is based on the actual frame populations established in March 2009 and in March 2010. The study variable in the survey TSSS is Monthly Turnover and exactly the same variable can be found (in retrospect) in the monthly Value Added Tax (VAT) returns. This means that we have values on the study variable for all enterprises in both frame populations.

Although this study employs the specific sample coordination PRN technique used at Statistics Sweden, similar PRN techniques are used for sample coordination in several countries. It is not unlikely that the significance and properties of the correlation obtained by these other methods would be quite similar to the correlation obtained by the method presented here. In Section 2, we describe the system used at Statistics Sweden for coordinated frame development and sample selection (the SAMU System). Section 3 presents background information on the TSSS. We present the formulae used for variance and correlation estimation in Section 4. Section 5 presents the simulation study. We conclude in Section 6 with general comments and ideas for future research.

2. The SAMU System

Statistics Sweden uses the SAMU system (Ohlsson 1995, Lindblom 2003) for the coordination of frame populations and sample selection from the BR. The SAMU system has three main objectives: (1) to obtain statistics comparable both in time and between surveys; (2) to ensure high precision in estimates of change over time; and (3) to spread the response burden between the businesses.

The SAMU utilizes a very clever and simple method of drawing coordinated samples. A random number, independently selected from a set of random numbers uniformly distributed over the interval (0, 1), is assigned to every new unit as it enters the BR, and the unit retains this value as long as it remains in the BR. "Closed-down" units (deaths) are deleted from the BR along with their random numbers. After the random number assignment is complete, the entire frame population is ordered by strata, with units in each stratum sorted in ascending random number sequence, and the first n_h units in the strata are selected. Ohlsson (1992) formally proves that the sampling technique used in SAMU is equivalent to STSI without replacement.

The sample coordination in SAMU introduces a dependence between the realized samples that would not be present if new independent random numbers were assigned to each unit on the updated frame prior to sample selection. Since the random number assignment by SAMU is permanent, the same random number is used in each subsequent sample selection after its initial assignment. Each new STSI is drawn using these permanent random numbers. In this way, the STSI incorporates the most recent changes from the updated BR. Furthermore, a large overlap with the most recent sample can be expected since a persistent unit has the same random number on both occasions. All current surveys benefit from frame populations stemming from the same updated version of the BR. However, a drawback of this is that the precision in estimates of change will be sensitive to the proportion of units that change stratum between sample occasions. On the other hand, the use of the latest updated version of the BR is also important, especially for the level estimates.

3. Background on the Turnover in the Service Sector Survey (TSSS)

The "Turnover in the Service Sector" survey (TSSS) conducted by Statistics Sweden produces detailed monthly and quarterly estimates of turnover changes in 138 domains according to the Statistical Classification of Economic Activities in the European Community (NACE Rev. 2). Monthly Turnover is the only variable collected in this survey, and change estimates – not levels – are published.

The year-to-year change in turnover, $\hat{t}_{m1}/\hat{t}_{m0}$, is an important published statistic, where $\hat{t}_{mt}$ is the combined (size strata within industry) ratio estimate (Särndal et al. 1992) of the turnover level for month (m) year ($t = 0$ or $t = 1$, with 1 as the most recent year). The auxiliary information used is annual turnover, the same information used for cut-off and in the stratification (see below). Large enterprises (selected with certainty) are excluded from the combined ratio estimator due to their large impact on the estimates. Their turnover sum is added to the combined ratio estimates (each of the minimal number of nonresponding large enterprises are individually imputed).

The survey covers the following industries, classified into the service sector according to NACE Rev. 2: Motor, wholesale and retail trade (45–47); Transportation and storage (49–53); Accommodation and food service activities (55–56); IT and real estate businesses (58–75); Administrative and support service (77–82); Education, human health and social work (85–88); Art, entertainment and recreation (90–96). NACE is derived from the International Standard Industrial Classification of all Economic Activities (ISIC).

The frame population for the TSSS consists of all active enterprises in the BR classified into the service sector according to the above definition. Annual turnover is used as a unit size measure in TSSS and enterprise level information on monthly turnover is collected from monthly VAT returns. The variable Annual Turnover is defined in TSSS as the sum of monthly turnover for the most recent 12-month period available at the sampling occasion. A cut-off limit is used in the survey, so that enterprises with an Annual Turnover less than 200,000 SEK (about $ 30,000) are excluded from the frame population and the samples. The final frame population consists of about 300,000 enterprises.

The stratification divides the frame population into 138 industrial strata based on economic activity. This stratification accommodates specialized domains of study as much as possible. Each industrial stratum is further subdivided into five size strata, with Annual Turnover as the unit size measure. Within each industry, one size stratum includes the largest enterprises, which are completely enumerated (a certainty or "take-all" stratum). The remaining units are grouped into four strata using the $cum\sqrt{f}$ method to determine stratum boundaries (Dalenius and Hodges 1959). Sample sizes in each stratum are obtained via optimum allocation (Neyman 1934), with Annual Turnover as the allocation variable. The total sample consists of about 12,000 enterprises. Approximately 2,500 enterprises are completely enumerated. These completely enumerated enterprises account for approximately 50 percent of the total turnover in the frame population.

Once a year, in March, a new frame population is established, and a new STSI is drawn using the SAMU. The frame population established and the sample drawn in March of a given year (t) are used for the period April year (t) to March year ($t+1$).

4. Variance and Correlation for Estimates of Change

4.1. Variance Estimation

As mentioned in Section 1, the complete frame population data is available for our study. Therefore, we can directly obtain the variances of the Monthly Turnover estimates at times $m0$ and $m1$ ($V(\hat{t}_{m0})$ and $V(\hat{t}_{m1})$, respectively) using the sampling formula variances for a STSI sample. The theoretical variance for the change estimate of Monthly Turnover is approximated by the Taylor Linearization formula:

$$V\left(\frac{\hat{t}_{m1}}{\hat{t}_{m0}}\right) \approx \left(\frac{t_{m1}}{t_{m0}}\right)^2 \left[\frac{V(\hat{t}_{m1})}{t_{m1}^2} + \frac{V(\hat{t}_{m0})}{t_{m0}^2} - 2\frac{C(\hat{t}_{m0},\hat{t}_{m1})}{t_{m0}t_{m1}}\right] \tag{1}$$

This is equivalent to:

$$V\left(\frac{\hat{t}_{m1}}{\hat{t}_{m0}}\right) \approx \left(\frac{t_{m1}}{t_{m0}}\right)^2 \left[\frac{V(\hat{t}_{m1})}{t_{m1}^2} + \frac{V(\hat{t}_{m0})}{t_{m0}^2} - 2\frac{\rho(\hat{t}_{m0},\hat{t}_{m1})\sqrt{V(\hat{t}_{m0})V(\hat{t}_{m1})}}{t_{m0}t_{m1}}\right] \tag{2}$$

However, a theoretical expression of $\rho(\hat{t}_{m0},\hat{t}_{m1})$ in Formula 2 would require the generation of all possible outcomes of pairwise coordinated samples from the two frame populations, and would require prohibitive computational resources.

4.2. Covariance/Correlation Estimation

The sample coordination method employed by SAMU makes estimating the correlation between the level estimates quite complicated because the size of the overlap between two samples is stochastic. Nordberg (2000) presents a complete and workable method for estimating this correlation under the SAMU sampling scheme. Related approaches can be found in Tam (1984), Laniel (1988), Hidiroglou et al. (1995), Berger (2004) and Wood (2008); Garås (1989) summarizes the preceding work on this approach conducted at Statistics Sweden.

Nordberg's method works when only sample data from each time period are available, as well as when values on the study variables are available for the whole frame population. For our study, we estimated the correlation by straightforward simulation and used those estimates as input to the analysis. In addition, we obtained correlation estimates using the method proposed by Nordberg (when values on the study variable are available for the whole frame populations). It is very useful for Statistics Sweden to compare the empirical measures to those obtained using Nordberg's method. This comparison (evaluation) confirms that Nordberg's method gives unbiased estimates. Comparison statistics for these empirical measures to those obtained using Nordberg's method are available upon demand.

Note that in this study the year-to-year change in turnover is based on the Horvitz-Thompson (HT) estimator of the turnover level for month (m) year ($t = 0$ or $t = 1$) instead of the ratio estimator used in the actual TSSS. This study aims to analyze how different choices of stratification variable, number of strata, and study variable are related to the overlap correlation (i.e., $\rho(\hat{t}_{m0}, \hat{t}_{m1})$ in Formula (2)) and the variances of the estimates of change. Use of the HT estimator, instead of the ratio estimator, makes the analysis presented in Section 5 more transparent and avoids confounding. The ratio estimator would add another factor to consider in the analysis, namely the correlation between the study variable and the auxiliary variable, which is very high in the TSSS but could possibly be lower for another choice of study variable.

To obtain empirical estimates of the correlation between the level estimates $\hat{t}_{m0}$ and $\hat{t}_{m1}$ (for each domain) in our simulation, we independently selected 10,000 coordinated samples from the frame population. Recall that the true variances of $\hat{t}_{m0}$ and $\hat{t}_{m1}$($V(\hat{t}_{m0})$ and $V(\hat{t}_{m1})$) are known. Let K = the number of generated pairwise coordinated samples (i.e., the number of replicates) selected in the simulation study ($k = 1, 2, \ldots,$ K). We obtained empirical sample-based estimates of variance and covariance as

$$\hat{V}(\hat{t}_{mt}) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{t}_{mtk} - \hat{\bar{t}}_{mt})^2, \tag{3}$$

$$\hat{C}(\hat{t}_{m0}, \hat{t}_{m1}) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{t}_{m0k} - \hat{\bar{t}}_{m0})(\hat{t}_{m1k} - \hat{\bar{t}}_{m1}) \tag{4}$$

where $t = 0$ or 1 and $\hat{\bar{t}}_{mt}$ is the average level estimate over the K samples.

We verified that 10,000 was a sufficient number of replicates by comparing the sample-based values of $\hat{V}(\hat{t}_{m0})$ and $\hat{V}(\hat{t}_{m1})$ to $V(\hat{t}_{m0})$ and $V(\hat{t}_{m1})$, respectively. The large number of replicates yielded variance estimates that were essentially unbiased over repeated samples, implying that the estimated correlation ($\hat{\rho}$) was likewise unbiased for ρ. Table 1 compares the abovementioned $\hat{V}(\hat{t}_{m0})$ and $\hat{V}(\hat{t}_{m1})$ to $V(\hat{t}_{m0})$ and $V(\hat{t}_{m1})$ obtained by samples using the STSI sampling design selected from the frame population data of the TSSS.

In addition, the number of sufficient replicates was confirmed by comparing the difference in obtained correlation estimates after 100, 500, 1,000, and up to 10,000 replicates to validate that 10,000 replicates were sufficient to ensure convergence.

Using Formulas (3) and (4), we estimated $\rho(\hat{t}_{m0}, \hat{t}_{m1})$ in Formula (2) in each domain.

Table 1. Comparison between sample-based and theoretical variances obtained by the sampling design used in TSSS

NACE Industry	$V(\hat{t}_{m0})$	$\hat{V}(\hat{t}_{m0})$	$V(\hat{t}_{m1})$	$\hat{V}(\hat{t}_{m1})$
45	117,419	117,901	147,626	148,695
46	1,052,889	1,031,004	1,304,284	1,294,880
47	518,376	520,264	583,215	590,050

Note that unlike the variance estimates, whose higher domain level estimates can be obtained by aggregating the independent lower level domain estimates, the covariance estimates must be computed separately for the aggregate domain and for the separate lower level subdomains. The covariance estimates are based on information collected at two time points and are therefore affected by enterprises changing lower level subdomain between the two sample occasions.

5. The Simulation Study

5.1. Simulation Study Design

The actual frame populations established in March 2009 and in March 2010 for the TSSS provide the study data. The study variable is Monthly Turnover, obtained retrospectively for all enterprises from the monthly VAT returns (for the majority of the enterprises) or from the Annual Income Tax returns (for a minor portion of the enterprises). In the latter case, an estimated Monthly Turnover was produced by dividing Annual Turnover by twelve. Due to the timing of the VAT returns, it is not possible to use the turnover values from monthly VAT returns in the production of the survey statistics.

We compare the three different STSI sampling designs, ranging from highly detailed (numerous size strata) to a single noncertainty stratum with a very heterogeneous population:

1. Each industry stratum has four sampled size groups and one take-all stratum (4-size gr.). This is the current design of the TSSS.
2. Each industry stratum has three sampled size groups and one take-all stratum (3-size gr.)
3. Each industry stratum has one sampled size group and one take-all stratum (1-size gr.)

Each design was applied to the same frame populations, with industry as the first level stratification variable. After determining the take-all (certainty) units, the remaining units were stratified into four, three and one noncertainty strata by unit size strata within industry (depending on design) using the $cum\sqrt{f}$ rule. In the tables below, we label four, three, and one noncertainty strata designs as "4-size gr.," "3-size-gr.," and "1-size gr."

Besides varying the number of strata, we considered the effects of alternative second level stratification variables (unit size variables) on the estimated precision. With the TSSS, the correlation between the stratification variable (Annual Turnover) and the study variable (Monthly Turnover) is very high. To extend the results to a less "ideal" situation – that is, reducing the correlation between the size measure and the study variable/s – we

restratified the frame populations using Number of Employees (collected from the BR) as a size measure and repeated the experiment (Note: although less correlated with Monthly Turnover, Number of Employees is a much more stable variable compared to Annual Turnover). In addition, we consider two study variables: Monthly Turnover and Annual Value Added. The study variable Annual Value Added is obtained retrospectively for all enterprises from the Annual Income Tax return. For both stratifications, optimum allocation based on Annual Turnover was used to determine the sample sizes in each stratum under the constraints that total sample size on the three-digit NACE level should be almost the same in all designs.

The estimates were produced at the two- and three-digit NACE Rev. 2 levels. To ensure comparability between the three different sampling designs, all designs have approximately the same sample size for each year in each three-digit NACE domain. Unfortunately, it was too time consuming to include all industries covered by the survey. Consequently, we restricted the analysis to a subset of the TSSS industries: Motor Trade (45), Wholesale Trade (46) and Retail Trade (47). These industries comprise about 75,000 enterprises and were chosen for their importance in the TSSS.

We selected independent pairwise samples per design from the 2009 and 2010 frames, replicating the SAMU PRN-coordination sampling procedure 10,000 times. For each replicate k, we generated a unique seed as the integer part of a random number uniformly distributed over the interval (0, 1) using the SAS RANUNI function (Fishman and Moore 1982), multiplied by a million. The replicate seeds were used to generate the permanent random numbers assigned to all enterprises in the frame population at time 0 (2009) and to the new enterprises in the frame population at time 1 (2010).

Tables 2a and 2b present aggregated information, from each stratification, on the number of enterprises in the frame populations, the number of enterprises in the samples (take-all and sampled), along with aggregated information on frame population overlap and sample overlap (averaging over repeated samples). The counts in the Overlap columns exclude take-all units as well as strata whose frame populations contain one common enterprise in the two years.

Since different variables are used for the two stratifications, the sets of take-all enterprises presented in Tables 2a and 2b do not coincide entirely. However, the difference between the two sets is very slight because an enterprise with large turnover usually has a large number of employees.

5.2. *Results*

We conducted all analyses on both the two- and three-digit NACE Rev. 2 levels. To save space, only the two-digit level results are included; however, the results on the three-digit level support the results on the two-digit level. Tables 3a and 3b show the gain in efficiency in terms of variance reduction (in percent) for the two-digit level change estimates, comparing the variance estimates obtained by using dependent SAMU samples ($VDep$) to the corresponding variance estimates obtained by using independent samples ($VInd$) with gain measured by $100 \bullet \left(1 - \frac{VDep}{VInd}\right)$.

The efficiency gained by using dependent SAMU samples rather than independent samples is quite substantial. At a minimum, a variance reduction of at least about

Table 2a. Sample Design Characteristics with Annual Turnover as Stratification Variable

				Realized Sample Sizes					
		Population		2009		2010			
Design	NACE industry	2009	2010	Take-all	Sampled	Take-all	Sampled	Overlap in population	Overlap in sample
4-size gr.	45	12,868	12,710	207	405	197	403	11,051	326
	46	26,855	26,366	717	783	664	790	22,481	584
	47	34,945	34,132	515	1,133	526	1,139	29,116	818
3-size gr.	45	12,868	12,710	207	402	197	401	11,064	331
	46	26,855	26,366	717	797	664	790	22,490	581
	47	34,945	34,132	515	1,103	526	1,102	29,129	816
1-size gr.	45	12,868	12,710	207	400	197	400	11,076	342
	46	26,855	26,366	717	800	664	802	22,593	673
	47	34,945	34,132	515	1,087	526	1,090	29,227	868

Table 2b. Sample Design Characteristics with Number of Employees as Stratification Variable

				Realized Sample Sizes					
		Population		2009		2010			
Design	NACE industry	2009	2010	Take-all	Sampled	Take-all	Sampled	Overlap in population	Overlap in sample
4-size gr.	45	12,868	12,710	192	418	183	418	11,078	332
	46	26,855	26,366	535	947	520	947	22,664	745
	47	34,945	34,132	477	1,149	494	1,128	29,190	818
3-size gr.	45	12,868	12,710	192	419	183	419	11,094	342
	46	26,855	26,366	535	945	520	945	22,692	762
	47	34,945	34,132	477	1,150	494	1,132	29,208	856
1-size gr.	45	12,868	12,710	192	418	183	427	11,090	359
	46	26,855	26,366	535	945	520	960	22,718	799
	47	34,945	34,132	477	1,147	494	1,130	29,265	891

Table 3a. Stratification by Annual Turnover

NACE industry	Measure Monthly Turnover			Measure Annual Value Added		
	4-size gr. Gain	*3-size gr. Gain*	*1-size gr. Gain*	*4-size gr. Gain*	*3-size gr. Gain*	*1-size gr. Gain*
45	22.3%	30.7%	74.1%	38.0%	42.7%	80.2%
46	24.6%	32.5%	66.6%	41.9%	48.7%	71.8%
47	36.3%	47.0%	78.1%	52.1%	57.1%	75.0%

Table 3b. Stratification by Number of Employees

NACE industry	Measure Monthly Turnover			Measure Annual Value Added		
	4-size gr. Gain	*3-size gr. Gain*	*1-size gr. Gain*	*4-size gr. Gain*	*3-size gr. Gain*	*1-size gr. Gain*
45	54.4%	63.1%	81.8%	63.8%	71.2%	82.0%
46	69.7%	74.3%	73.0%	56.1%	59.4%	71.7%
47	62.6%	67.9%	80.2%	55.9%	65.3%	78.3%

20 percent is attained with the highly stratified design (4-size gr). As the number of strata decreases, the efficiency gains from the dependent SAMU samples are more evident. The gains in efficiency are especially noticeable when the stratification and study variables are less strongly correlated (Table 3b), although the gain is not negligible when the stratification and study variables are highly correlated (Table 3a).

Tables 4a through 4d present the standard errors of the change estimates in percentage points for each sampling design. *SEDep* is the standard error obtained by using overlapping SAMU samples, *SEInd* is the standard error obtained by using independent samples and *Corr* $\left(\hat{\rho}(\hat{t}_{m0}, \hat{t}_{m1})\right)$ is the estimated overlap correlation obtained using SAMU samples.

For the majority, the most detailed stratification (4-size gr.) yields the smallest *SEDep*. In general, the improvements in precision for the input level (total estimates) offset the smaller sample overlap compared to the other design. The magnitude of the overlap correlation increases as the number of size groups (strata) decreases. The difference in precision with four and three size groups (noncertainty strata) is small for *SEDep*, compared to the difference in precision with three and one size groups in many cases. Often, the increase in *SEInd* caused by reducing the number of size groups from four to three is offset by the increased overlap correlation, and there is no detrimental effect on the precision of the estimate of change. However, when only one size group is employed, both the *Corr* and *SEInd* increase substantially, and the increased overlap correlation cannot compensate for the increased *SEInd*.

By comparing corresponding cells in Tables 4a and 4b and in Tables 4c and 4d, we can examine the relationship between the stratification variable and the study variable on the overlap correlation. The results presented in Tables 4a and 4b show that the overlap correlation of Monthly Turnover increases substantially when Number of Employees is the stratification variable. This increase is probably a function of the stability of Number of Employees in contrast to the more volatile Annual Turnover. Because the Number of

Table 4a. Stratification by Annual Turnover, Monthly Turnover Measured

NACE industry	Four sampled size groups			Three sampled size groups			One sampled size group		
	SEDep	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*
45	2.2%	2.5%	0.22	2.5%	3.0%	0.31	4.1%	8.0%	0.74
46	1.4%	1.6%	0.25	1.5%	1.8%	0.32	3.1%	5.3%	0.67
47	1.7%	2.1%	0.36	1.7%	2.3%	0.47	2.7%	5.8%	0.78

Table 4b. Stratification by Number of Employees, Monthly Turnover Measured

NACE industry	Four sampled size groups			Three sampled size groups			One sampled size group		
	SEDep	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*
45	6.8%	10.1%	0.55	6.7%	11.1%	0.63	6.3%	14.7%	0.83
46	4.8%	8.7%	0.71	4.8%	9.5%	0.75	6.2%	11.9%	0.75
47	1.9%	3.1%	0.63	1.8%	3.2%	0.68	2.5%	5.6%	0.80

Table 4c. Stratification by Annual Turnover, Annual Value Added Measured

NACE industry	Four sampled size groups			Three sampled size groups			One sampled size group		
	SEDep	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*
45	4.3%	5.4%	0.40	4.3%	5.7%	0.45	5.0%	11.2%	0.80
46	2.8%	3.6%	0.42	2.7%	3.8%	0.49	4.2%	7.8%	0.72
47	1.9%	2.7%	0.52	1.9%	2.9%	0.57	2.9%	5.8%	0.75

Table 4d. Stratification by Number of Employees, Annual Value Added Measured

NACE industry	Four sampled size groups			Three sampled size groups			One sampled size group		
	SEDep	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*	*SEDep*	*SEInd*	*Corr*
45	4.0%	6.6%	0.64	4.4%	8.2%	0.71	4.8%	11.4%	0.82
46	3.1%	4.7%	0.57	3.4%	5.4%	0.60	4.4%	8.3%	0.72
47	1.7%	2.6%	0.56	1.7%	3.0%	0.65	2.4%	5.1%	0.78

Employees in an enterprise tends to remain constant, the enterprise is often retained in the same stratum in consecutive sampling occasions, facilitating larger sample overlap. Although the correlation due to overlap is higher when obtained with the more stable stratification variable, this does not imply that the change estimates are likewise more precise. The correlations presented in Table 4a are consistently lower than their Table 4b counterparts, but the *SEDep* estimates are also considerably lower. Recall that the stratification variable and study variable used in Table 4a are very highly correlated, whereas the stratification and study variables used in Table 4b are not. In the former case, the variance estimates of monthly turnover (*SEInd*) are much lower than those obtained

using the other stratification. The increased *Corr* due to a larger overlap does not compensate for the larger variance estimates of the level estimates.

Tables 4c and 4d demonstrate similar patterns with a different study variable (Annual Value Added). Here, the overlap *Corr* increases as the number of strata decreases. As in Tables 4a and 4b, using Number of Employees as a stratification variable again increases the magnitude of the overlap *Corr*. Again, the differences in precision (*SEDep*) obtained between three and four size group stratifications are very small. Finally, the increased *Corr* due to the large overlap in the one sampled size group design largely compensates for the increased variance of the level estimates, although overall precision still tends to be lower than with the more stratified designs. The comparisons of the *Corr* between the designs with different stratification variables may be somewhat confounded by the different size measures. Recall that there are slightly different sets of take-all enterprises for both designs, which in turn affects the sampling variance.

Finally, we compare corresponding *Corr* values in Tables 4a to 4c and Tables 4b to 4d. The results in Tables 4a and 4c are based on exactly the same sampling design; the only difference is that Monthly Turnover is replaced by Annual Value Added as study variable. A comparison between Tables 4a and 4c reveals that the realized values of *Corr* are very close when the sampling design employs one sampling strata (One Sampled Size Group). In this case, the effect of stratification variable and size of sample overlap is eliminated and the only difference is due to different study variables. This indicates that the correlation between two Annual Value Added values, observed on the same unit at two different occasions, have similar patterns as those seen with Monthly Turnover when the stratification is not very detailed. However, the amount of realized *Corr* increases substantially when four (and three) sampled size groups are used (regardless of stratification variable) when Monthly Turnover is replaced by Annual Value Added as the study variable. We suspect that this phenomenon is related to size of sample overlap and the correlation between stratification and study variables.

6. Conclusions and Future Research

In this article, we present a study that examines the effects of degree of stratification, correlation between stratification variables and study variables, and overlap correlation between the level estimates $\hat{t}_{m0}$ and $\hat{t}_{m1}$ obtained by using the PRN technique utilized at Statistics Sweden on the precision of estimates of change (level estimates produced by the HT-estimator). The studied SAMU method is easy to implement, but the sample designers have to make many decisions. Specifically, they must balance the need for highly stratified designs – which reduce the variance of the level estimates – with the need for a substantive sample overlap to increase the correlation between the adjacent level estimates to increase the precision of the change estimates (the primary statistics of interest).

One conclusion from the study is that the overlap correlation is of less importance for the precision in estimates of change over time when study variable and stratification variable are highly correlated. In this case, the precision in estimates of change benefits most from the high precision in each level estimate. When the correlation between the stratification variable and the study variable decreases or when a more stable stratification variable was used, such as Number of Employees, we found that using a moderately

stratified design (three noncertainty strata instead of four) with the overlapping SAMU samples created a sufficiently high correlation to offset the increase in level estimate variances.

Since the study variable in TSSS (Monthly Turnover) is known in retrospect for the whole frame population, it was possible to estimate the overlap correlation by simulation in this study. In most other surveys the study variable values would be available only from a *single* sample from each time period. The method proposed by Nordberg (2000) yields unbiased estimates of the correlation between the level estimates $\hat{t}_{m0}$ and $\hat{t}_{m1}$ obtained by overlapping SAMU samples. However, if the proportion of enterprises that change stratum between two sample occasions is substantial the correlation estimates can become quite variable. This is the case in the TSSS, where the stratification variable Annual Turnover is fairly volatile, causing enterprises to change stratum rather frequently. If Monthly Turnover from an earlier time period can be used as a proxy variable for Monthly Turnover for the actual time period, then the overlap correlation could be estimated in practice by the same simulation method as used in the present study. Examining the effect of this procedure will be an issue for further study. Another important question for future study is the effect on the overlap correlation occurring when different survey designs, as well as different estimators, use the SAMU PRN-coordination method.

7. References

Berger, Y. 2004. "Variance Estimation for Measures of Change in Probability Sampling." *The Canadian Journal of Statistics* 32: 451–467. DOI: http://dx.doi.org/10.2307/3316027.

Dalenius, T. and J.L. Hodges. 1959. "Minimum Variance Stratification." *Journal of the American Statistical Association* 54: 88–101. DOI: http://dx.doi.org/10.2307/2282141.

Fishman, G.S. and L.R. Moore. 1982. "A Statistical Evaluation of Multiplicative Congruential Random Number Generators with Modulus $2^{31}-1$." *Journal of the American Statistical Association* 77: 29–136. DOI: http://dx.doi.org/10.1080/01621459.1982.10477775.

Garås, T. 1989. *Estimators of Change in Dynamic Populations*. Memo: Statistics Sweden. (In Swedish).

Hidiroglou, M., C.-E. Särndal, and D. Binder. 1995. "Weighting and Estimation in Business Surveys." In *Business Survey Methods*, edited by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge, and P.S. Scott, 477–502. New York: John Wiley & Sons.

Laniel, N. 1988. "Variances for a Rotating Sample from a Changing Population." In *Proceedings of the Business and Economic Statistics Section: American Statistical Association*. 246–250.

Lindblom, A. 2003. *SAMU – The System for Coordination of Frame Populations and Samples from the Business Register at Statistics Sweden*. Background Facts on Economic Statistics 2003:3, Statistics Sweden. Available at: http://www.scb.se/statistik/OV/AA9999/2003M00/X100ST0303.pdf (accessed September 1, 2014).

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method Stratified Sampling and the Method of Purposive Selection." *Journal of Royal Statistical Society* 97: 558–606. DOI: http://dx.doi.org/10.2307/2342192.

Nordberg, L. 2000. "On Variance Estimation for Measures of Change When Samples are Coordinated by the Use of Permanent Random Numbers." *Journal of Official Statistics* 16: 363–378. Available at: http://www.jos.nu/Articles/abstract.asp?article=164363 (accessed September 1, 2014).

Ohlsson, E. 1992. *SAMU – The System for Co-ordination of Samples from the Business Register at Statistics Sweden*. R&D Report, Statistics Sweden, 1992:18.

Ohlsson, E. 1995. "Coordination of Samples using Permanent Random Numbers." In *Business Survey Methods*, edited by B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge, and P. Kott, 153–169. New York: John Wiley.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Tam, S.M. 1984. "On Covariances from Overlapping Samples." *The American Statistician* 38: 288–292. DOI: http://dx.doi.org/10.1080/00031305.1984.10483227.

Wood, J. 2008. "On the Covariance Between Related Horvitz-Thompson Estimators." *Journal of Official Statistics* 24: 53–78. Available at: http://www.jos.nu/Articles/abstract.asp?article=241053 (accessed September 1, 2014).

Received December 2012
Revised August 2014
Accepted August 2014