Use of a delayed sampling design in business surveys coordinated by permanent random numbers

Fredrik Jonsson, Eva Elvers, Jörgen Brewitz[1]

Abstract[2]

Sample coordination is used to increase, or decrease, the overlap between samples over time and between surveys. Permanent random numbers (PRN's) is one method to achieve coordination, balancing statistical accuracy and the distribution of response burden. Updating registers and frames with data from samples coordinated by PRN's may lead to bias in estimates, as drawn samples become more updated than the frame and population as a whole. In order to reduce the impact of feedback bias we consider the possibility of introducing delayed sampling designs, in the sense that corrections of randomly selected units should not affect sampling probabilities until the corresponding PRN's have been rotated. The implications of the approach are investigated, in terms of the contribution to the total survey error as compared to an unlimited use of feedback in sampling and estimation.

Key Words: Coordinated sampling, business registers, permanent random numbers, feedback bias

1. Introduction

Business statistics provides many challenges. A statistical office runs a set of annual, quarterly, and monthly surveys on business activities, the labor market, trade, prices, etc. The statistics should be coherent, i.e. the surveys should use the same definitions and the same methods.

Estimation of change may be an important part of a survey. Changes over time are due both to population changes and changes in values of variables. There are advantages with overlap between samples from an accuracy point of view. Considering response burden, it may be desirable for a business to be "in" for a period of time and then "out" for another. Sample coordination may be used to increase, or decrease, the overlap between samples over time and between surveys. Permanent random numbers (PRN's) is one method to achieve such coordination, balancing accuracy and the distribution of response burden (Srinath & Carpenter, 1995), (Ohlsson, 1995), (Kalton, 2009).

The Business Register is a common basis for business statistics. Maintaining the business register is essential to capture births, deaths, splits, changes in activity, and contact information, as far as possible, before creating a common frame for the surveys. Several sources are used for updates, e.g. administrative systems held by tax offices and annual or sub-annual register inquiries. Frames are created from the business register, often a "frozen" version of the business register (Colledge, 1995), (Eurostat, 2010, p. 18), (Smith, 2013, p. 167).

---

[1] Fredrik Jonsson, Statistics Sweden, email: fredrik.jonsson@scb.se. Eva Elvers, email: eva.elvers@scb.se. Jörgen Brewitz, email: jorgen.brewitz@ scb.se.

[2] A preliminary version of this paper, entitled *Handling survey feedback in business statistics*, was presented at the ICES-V conference.

Updating registers and frames with data from samples in repeated surveys is problematic if there is a controlled overlap of samples, as it may lead to bias in estimates. For instance, information about deaths may be quicker from the survey than from the regular sources. If such information is fed into the register and the new frame, the next sample that is drawn will be more updated than the frame and population as a whole (Colledge, 1995), (Ohlsson, 1995), (Hedlin & Wang, 2004), (Hidiroglou & Lavallée, 2009), (Eurostat, 2010, p. 161), (Smith, 2013, p. 174), (ONS, 2001, p. 59).

Different approaches for reducing the impact of feedback bias have been suggested. The strategies can be classified as being either preventive or adjusting. For the effect of removing ineligible units, (Hedlin & Wang, 2004) proposed an adjusted estimation procedure, based on consistent estimates of the number of eligible units in the population. A simple preventive approach is to avoid updating the business register from statistically dependent sources (Ohlsson, 1995), (ONS, 2001, p. 63).

Aiming at a more refined preventive approach, we consider introducing *delayed sampling designs*, in the limited sense that corrections of randomly selected units should not affect sampling probabilities until the corresponding PRN's have been rotated. The aim of this paper is to investigate the introduction, motivation and implications of this strategy. In particular, we try to answer: *a) What is the contribution of feedback bias to the total survey error and how should it be assessed? b) What is the bias/variance trade-off for introducing a delayed sampling design and how should it be evaluated?*

2. Delayed sampling designs in practice

A definition of a delayed sampling design is presented in Section 3.2. In the following sections we give a brief outline of an approximate implementation. Some simplification is suggested, in the sense that the conditions of Section 3.2 are not strictly fulfilled, aiming at a balance between feedback bias (Section 4), costs, and other error sources (Sections 3.1 and 3.3).

2.1 Delayed sampling frames

We assume that a business register is maintained and regularly updated, with the aim of describing current activity and providing contact information for a business population. A set of variables describe each business unit. The variables can be characterized as either primary or secondary (a secondary variable can be deduced from primary variables). Each primary variable is complemented by a source and a time stamp, describing how the information was obtained, by referring e.g. to an administrative source, a register inquiry, or a repeated survey.

A *regular sampling frame* is obtained from the prioritized descriptions of the business population provided by the business register at a given point in time. Moreover, permanent random numbers (PRN's) are associated to each business unit, complemented by corresponding time stamps. (Smith, 2013)

A *delayed sampling frame* is characterized by the requirement that recent updates of register variables with source stamps referring to repeated surveys are not allowed. In constructing the delayed sampling frame, we distinguish primary and secondary variables. For primary variables, we distinguish updates from a repeated survey, such that

the time stamp of the update is more recent than the time stamp of the corresponding PRN in the regular sampling frame. Such updates need to be replaced by the preceding value in the business register. Then, secondary variables can be deduced from primary variables, using the deduction rules of the regular sampling frame.

2.2 Obtaining a stratified sampling frame

Considering a given survey, we now assume that a target population can be defined as a subset of the current business population, through a set of selection criteria operating on business register variables. A corresponding frame population is to be derived and divided into sampling strata.

We first select business units in take-all strata, through selection criteria applied to the regular sampling frame. Similarly, business units in take-none strata are distinguished from the regular sampling frame. Finally, take-all and take-none strata need to be complemented by take-some strata. The selection is made with rules operating on the delayed sampling frame, excluding all units already included in take-all and take-none strata.

Corrections of randomly selected units may still have a limited effect on sampling probabilities with this procedure, even before the corresponding PRN's have been rotated. Indeed, a business unit may first be selected in a take-some stratum, then corrected, and due to the correction selected in a take-all stratum at a subsequent sampling occasion.

2.3 Descriptions of business units

From the previous steps we obtained a frame population, divided into sampling strata. Next, frame units need to be described in various dimensions. The descriptions serve multiple purposes: allocating sample sizes; assigning size measures for probability-proportional-to-size sampling; identifying over-coverage prior to data collection; providing auxiliary data for calibrated estimation or imputation methods; assigning an appropriate questionnaire; contact information; associating target domains; etcetera.

As a general rule we consider non-delayed frame data for describing business units, unless there is reason to believe in major distortive effects in calibrated estimation, assignment of sample sizes, or probability-proportional-to-size sampling.

3. Unbiased estimation

A general framework for PRN-dependent sampling design and sample selection is introduced in the following sections. We refer to the appendix for a proof of Propositions 1-3.

3.1 Random sampling designs

Consider a sampling frame F consisting of N units, and a probability space Ω represented by a uniform random variable U. The random variable U is introduced for the purpose of coordinated sample selection, and referred to as permanent random numbers, or PRN's.

A *random sampling design* refers to a PRN-dependent assignment of sampling probabilities for each subset of the sampling frame:

$$D(\omega) = \{D_s(\omega)\}_{s:\, s\subset F}, \qquad \sum_{s:\, s\subset F} D_s(\omega) = 1$$

An associated *sample selection procedure* is deduced from a partitioning of Ω into disjoint, possibly empty, subsets $\Omega_s(\omega)$ related to the sampling design through the following equations:

$$D_s(\omega) = P\big(U \in \Omega_s(\omega)\big), \qquad \forall s \subset F$$

Then, the unique sample s is selected for which U is an element of $\Omega_s(U)$:

$$I\{S = s\} = I\{U \in \Omega_s(U)\}$$

Definition 1. *The sample selection procedure S is consistent with the random sampling design D, whenever $P(S = s \mid D) = D_s$ for all possible subsets $s \subset F$.*

By definition, the sample selection procedure associated to a non-random sampling design is always consistent.

We refer to the following point and variance estimation procedure as the *Horwitz-Thompson estimator* corresponding to a random sampling design, a sample selection procedure, and a target variable y:

$$\hat{t} = \sum\nolimits_S \breve{y}_k\,, \qquad \hat{V}(\hat{t}) = \sum\sum\nolimits_S \frac{(\pi_{k,l} - \pi_k\pi_l)}{\pi_{k,l}} \breve{y}_k \breve{y}_l$$

Here, π_k and $\pi_{k,l}$ refer to first and second order (random) inclusion probabilities $\pi_k = \sum_{s:\, k\in s} D_s$, $\pi_{k,l} = \sum_{s:\, k,l\in s} D_s$, whereas $\breve{y}_k$ refers to probability-expanded target values $\breve{y}_k = y_k/\pi_k$.

Proposition 1. *The Horwitz-Thompson estimator corresponding to a random sampling design satisfying $P(\pi_k > 0) = 1$ for all $y_k \neq 0$, and a consistent sample selection procedure, is unbiased for $\sum_F y_k$. The associated variance estimator is also unbiased, provided $P(\pi_{k,l} > 0) = 1$ for all $y_k y_l \neq 0$.*

Thus, unbiased estimation is obtained, provided the inclusion probabilities for active target units are non-zero, with probability one. This condition could be violated, in particular when sampling feedback influences the variable constraints for the target population. Violations of the condition can be referred to as under-coverage for a random sampling design. Ideally, the target population for the sampling design should be sufficiently large and easy to cover, with no loss of target units due to uncorrected misclassifications.

Distribution of sampled units with respect to various domains of study is discussed in Section 3.3.

3.2 Separating the sampling design

Consider decomposing U into independent components $U_{j,k}$, $j = 1, \dots, M$, $k = 1, \dots, N$. For each sampling unit, we think of a set of M independent PRN's, to be used for sample selection in different blocks of surveys, and at different points in time[3].

Definition 2. *The sample selection procedure S is separated from the associated sampling design, whenever $U = (U_1, U_2)$ with U_1 and U_2 independent, and:*

$$\Omega_s(\omega) = \tilde{\Omega}_s(\omega_2) \times \Omega_2$$

With the above definition, we refer to U_1 as *primary PRN's*, and to U_2 as *secondary PRN's*, for the given sampling occasion. Note that the primary PRN's have a primary influence on the selection of sampled units:

$$I\{S = s\} = I\{U_1 \in \tilde{\Omega}_s(U_2)\}$$

whereas only the secondary PRN's have an influence on sampling frequencies:

$$D_s(\omega) = P\left(U_1 \in \tilde{\Omega}_s(\omega_2)\right)$$

Proposition 2. *A sample selection procedure separated from the associated sampling design is always consistent.*

A particular scenario occurs if the PRN's are replaced (or "rotated") at certain points in time. Introducing a time dimension, we obtain the following decomposition, with U_0 referring to all primary PRN's at the given point in time:

$$U = (U_0, U_{-1}, U_{-2}, \dots, U_{-t})$$

Definition 3. *A random sampling design is delayed if it is independent of all primary PRN's at the given point in time, i.e. only dependent on* $(U_{-1}, U_{-2}, \dots, U_{-t})$.

3.3 Identification of domains

For some level of detail corresponding to a set of domains (subsets of the sampling frame), consider estimating the domain totals

$$t_G = \sum\nolimits_G y_k\,, \qquad G \in \{G_1, \dots, G_L\}$$

by restricting the Horwitz-Thompson estimator (as described in Proposition 1) with a domain identifier Γ, leading to a so-called *domain estimator*:

$$\hat{t}_G = \sum\nolimits_{S \cap \{\Gamma_k = G\}} \breve{y}_k\,, \qquad \hat{V}(\hat{t}_G) = \sum\sum\nolimits_{S \cap \{\Gamma_k = G\}} \frac{(\pi_{k,l} - \pi_k \pi_l)}{\pi_{k,l}} \breve{y}_k \breve{y}_l$$

[3] This can be seen as an approximation of the system currently at use in Statistics Sweden, where PRN's are rotated at fixed time intervals, and different blocks of surveys use different start points and directions for sample selection at the PRN unit interval, cf. (Ohlsson, 1992).

Ideally, Γ identifies the domain G without any coverage error. Unbiased estimation then follows from Proposition 1.

Domain estimators only require a correct identification among sampled units; the identification of domains is not a priori a part of the sampling design. However, the ability to predict the relative error of the domain estimator is often important, in particular when determining the number of units to be sampled.

Assuming unbiased estimation in the case of stratified random sampling, we obtain from the above point and variance estimators the following expression for the estimated coefficient of variation:

$$\widehat{cv}_G^2 = \sum_{h=1}^{L} \tau_h^2 \frac{cv_h^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right)$$

Here, summation refers to all strata covering the target domain, τ_h to the proportion of the target value covered by the given stratum, n_h to the number of sampled target units within the given stratum, cv_h to the coefficient of variation of the target variable among sampled target units within the given stratum, and N_h to the total number of units within the given stratum.

There are several ways in which a prediction model may underestimate the variability of a domain estimator. In particular, a predicted coefficient of variation may:

(i) Fail to acknowledge the contribution from certain strata to the variability of the point estimate.
(ii) Overestimate the effective sample sizes.

For modest sampling frequencies, the effect of a distorted effective sample size on $\widehat{cv}_G^2$ can be approximated by the proportion N_{frame}/N_{true}, referring to the total number of business units in the frame and target population respectively. As to (i), the impact can potentially be more dramatic. *Outlier-effects* may occur if the corresponding business units are sampled at very low sampling frequencies (Beaumont & Rivest, 2009).

A different scenario takes place if the domain identifier Γ is constructed from register data, with independent corrections prior to the generation of the PRN's actively in use for the given survey, and/or independent corrections from non-coordinated surveys. Some coverage error may still remain:

$$CovErr_G = \sum_{\{\Gamma_k = G\}} y_k - \sum_G y_k$$

Proposition 3. *For a domain identifier independent of sample selection, the bias of the domain estimator equals the expected coverage error for the given target variable. The corresponding variance estimator is negatively biased by the variance of the coverage error.*

It is not unlikely that corrections of the domain identifier concentrate to certain survey actions. As a result, the remaining bias may have different characteristics in different blocks of surveys. This can be a disadvantage from the perspective of a coordinated

survey output. The effect is likely most notable for domains where the sampling frequency is low, so that the corrections have a negligible impact on the over-all quality of the sampling frame.

4. Feedback bias

Any bias which can be explained by having a sampling mechanism not independent of the feedback procedure could be referred to as *feedback bias* (Hedlin & Wang, 2004).

Estimators of feedback bias were proposed in (Hedlin & Wang, 2004), considering simple random sampling and the effect of removing ineligible sampling units from the sampling frame. A simulation study suggested that the removal of ineligible sampling units in a frame with 5% ineligible units can result in 2-3% bias.

The aim of Section 4.1 is to propose a simple measure for the contribution of feedback bias to the total survey error. The measure is compared to the true effect in Section 4.2, based on simulated data.

4.1 A simple measure

We assume that a coordinated sample is selected, consisting of a proportion ρ of previously non-sampled "rotated" units, and a proportion *1-*ρ of previously sampled "non-rotated" units. Moreover, we assume that a target group *G* can be identified within the selected sample. However, the identification of *G* within the sampling frame should be distinguished from corrections of coverage errors obtained during data collection. No coverage error remains within the subsample of previously sampled units, if frame imperfections are detected and fed back to update the business register and subsequent sampling frames. The resulting bias corresponding to the non-rotated part of the sampling frame can be described by the following result.

Proposition 4. *The Horwitz-Thompson estimator* $\hat{t} = (N_{frame}/n)\sum_S y_k$ *corresponding to a random sample S of n units among a population F of* N_{true} *units has a relative bias with respect to* $t = \sum_F y_k$ *given by the relative coverage error in numbers of business units.*

In practice, stratified random sampling is often used, and the relative coverage error may vary between sampling strata. In particular, a completely enumerated stratum does not contribute to the total feedback bias. Thus, we propose the following simple measure:

$$\frac{\hat{t}_G - t_G}{t_G} = (1-\tau)(1-\rho)\left(\frac{N_{frame} - N_{true}}{N_{true}}\right)$$

where τ refers to an estimate of the proportion of the target value covered by take-all strata, ρ to an estimate of the proportion of previously unsampled units in take-some strata, N_{frame} to the number of non-rotated units covered by take-some strata and associated to *G* within the sampling frame, and N_{true} to an estimated number of non-rotated units covered by take-some strata and included in *G*.

The relative coverage error in numbers of business units could be estimated based on an independent register inquiry, cf. e.g. (Hidiroglou & Lavallée, 2009, p. 446). A different

approach is to study historical corrections fed back by sample surveys to the business register over a 5 or 10 year period, and to compare with an estimate of the corresponding number of sampled units.

4.2 A simulation study

The aim of the following example is to consider a given business survey repeated over time, facing systematic misclassifications of economic activity. Attention was restricted to three areas of economic activity (see Table 2).

Table 1: Selected areas of economic activity

	Description
M71	Architectural and engineering activities; technical testing and analysis
F	Construction
D	Electricity, gas, steam and air conditioning supply

Based on reported economic activities in the Swedish business register at baseline year 2008, a *true* economic activity was randomly assigned to each business unit, according to fixed transition probabilities. For a typical realization of this process, consider Table 2. Clearly, a situation is simulated where approximately 10% of the units classified in M71 rather belong to F, whereas the remaining transition probabilities of erroneous classification are relatively small.

Table 2: Random assignment of correct classifications to business units according to fixed transition probabilities.

Reported	*Correct classification*			
	M71	*F*	*D*	*Total*
M71	8 594	1 010	3	9 607
F	49	23 989	2	24 040
D	1	1	139	141
Total	8 644	25 000	144	33 788

A sample-based yearly survey was considered, repeated over five consecutive years, estimating the total yearly turnover in each of the areas of economic activity, using simple random sampling cross-stratified according to three size classes based on number of employees and the three types of economic activity. Sample sizes within strata were assigned, considering the corresponding sampling rates of the structural business survey in Sweden. Business units assigned to take-all strata in the target survey were taken out of consideration. The five sampling occasions were coordinated by using the principles of the Swedish system for coordination of business surveys (SAMU), with a rotation rate of approximately 20% among smaller business units.

It was assumed that the true classification of each business unit could be detected when sampled (through the responses to the questionnaire). Three strategies were compared (summarized in Table 3), with differences as to whether correct classification was reported back to update the business register (and influence forthcoming sampling occasions), and as to whether the total turnover was distributed to each area of economic activity according to the observed corrected classifications.

Table 3: Summary of strategies for sampling and estimation.

Strategy	*Sampling*	*Estimation*
1	No feedback	No feedback
2	No feedback	Feedback
3	Feedback	Feedback

The resulting estimated values, when averaged over 1000 independent simulations (assigning PRN's at baseline year 2008 and true classifications of economic activity), is shown in Table 4.

Table 4: Expected estimated value divided by true value multiplied by 100 at year 2012.

	Strategy 1	*Strategy 2*	*Strategy 3*
M71	105	100	104
F	98	100	99
D	110	100	103
Total	100	100	100

There is a notable amount of feedback bias for the full use of feedback in Strategy 3. For instance, the estimate for M71 is suffering from the removal of ineligible sampling units from the corresponding sampling strata.

The true value for M71 is overestimated by 4% at year 2012, to be compared with 10% ineligible units in the sampling frame at year 2008. The indicative measure in Section 4.1 suggests a positive bias:

$$1 * 0.8 * 0.10 = 0.08$$

A reason to the difference between 4% and 8% might be that several size classes contribute to the estimate for M71 in a stratified design, and that sampling rates are higher in sampling strata corresponding to larger enterprises. Thus, a considerable amount of all misclassifications among larger enterprises might be fully corrected at year 2012 for the simulated model, so that 10% is overestimating the proportion of misclassifications in the sampling frame at the given sampling occasion.

5. Conclusions

The motivation for implementing a strategy avoiding feedback bias should arise from an evaluation of the contribution of feedback bias to the total survey error, currently and in the future. A simple measure was proposed in Section 4.1, and evaluated in Section 4.3.

We have highlighted three potential problems with the use of a delayed sampling design:

a) A loss of target units normally included in the frame population through sampling feedback (cf. Section 3.1, *under-coverage*).
b) Non-representative sampling weights for representative observations (cf. Section 3.3, *outlier-effects*).
c) Uncertainties in the effective sample size when estimating the target value for a given domain (cf. Section 3.3, *effective sample size*).

Problems a)-c) should be compared with the feedback bias associated with an unrestricted use of sampling feedback. Problems a) and b) may increase with a delayed sampling design, depending on the sensitivity of the sampling design to existing frame imperfections. Problem c) is comparable to feedback bias, as it relates proportionally to the relative amount of misclassifications fed back by survey sources for a given domain (Sections 3.3 and 4.2). Moreover, the analysis suggests that a delayed sampling design is superior to feedback bias, except for estimation at very large expected relative errors.

One may argue that corrections of frame imperfections observed during data collection should be used to as large extent as possible in updating the business register, and in subsequent survey design. Meanwhile, it is important to maintain the coherence between various survey outputs. This requires appropriate methodology in survey design, but also well-functioning feedback routines between business surveys and the business register.

Appendix

Proof of Proposition 1. By the assumption $\pi_k > 0$ for all $y_k \neq 0$, we may express the estimator as:

$$\hat{t} = \sum\nolimits_F y_k I_k/\pi_k\,, \qquad I_k = \sum\nolimits_{s:\, k\in s} I\{S = s\}$$

Using the assumption of a sample selection procedure consistent with the sampling design,

$$E(I_k/\pi_k \mid D) = \frac{1}{\pi_k} E(I_k \mid D) = \frac{1}{\pi_k} \sum\nolimits_{s:\, k\in s} P(S = s \mid D) = 1$$

By conditional expectation, we conclude that $E(\hat{t}) = \sum_F y_k$. For the variance estimator, using the above expression for $\hat{t}$:

$$V(\hat{t}) = \sum\nolimits_F \sum\nolimits_F y_k y_l \big(E\big(\check{I}_k \check{I}_l\big) - E\big(\check{I}_k\big)E\big(\check{I}_l\big)\big)$$

Also, since $\pi_{k,l} > 0$ for all $y_k y_l \neq 0$

$$E\left(\hat{V}(\hat{t})\right) = \sum\nolimits_F \sum\nolimits_F y_k y_l \big(E\big(\check{I}_k \check{I}_l\big) - E(I_k I_l/\pi_{k,l})\big)$$

Hence, to verify that $E\left(\hat{V}(\hat{t})\right) = V(\hat{t})$ it suffices to show that $E\big(\check{I}_k\big) = 1$ and that $E(I_k I_l/\pi_{k,l}) = 1$. Now, $E\big(\check{I}_k\big) = 1$ was an implicit part of the proof for unbiasedness of $\hat{t}$. Similarly,

$$E(I_k I_l/\pi_{k,l} \mid D) = \frac{1}{\pi_{k,l}} E(I_k I_l \mid D) = \frac{1}{\pi_{k,l}} \sum\nolimits_{s:\, k,l\in s} P(S = s \mid D) = 1$$

Proof of Proposition 2. Since D only depends on ω_2, by definition of conditional expectation, it suffices to verify that, for any $A \subseteq \Omega_2$,

$$\int D_s(\omega_2) I\{\omega_2 \in A\}\ dP_2(\omega_2) = P(\{U \in \Omega_s(U)\} \cap \{U_2 \in A\})$$

Now, $U \in \Omega_s(U)$ is equivalent to $U_1 \in \tilde{\Omega}_s(U_2)$ according to the assumption. Due to independence of U_1 and U_2, the right hand side of the above equation can thus be expressed as:

$$\int \int I\{\omega_1 \in \tilde{\Omega}_s(\omega_2)\} I\{\omega_2 \in A\}\ dP_1(\omega_1) dP_2(\omega_2)$$
$$= \int P(U_1 \in \tilde{\Omega}_s(\omega_2)) I\{\omega_2 \in A\}\ dP_2(\omega_2)$$

Finally, by definition of D_s,

$$D_s(\omega_2) = P\big(U \in \Omega_s(\omega_2)\big) = P(U_1 \in \tilde{\Omega}_s(\omega_2))$$

Proof of Proposition 3. Extending the proof of Proposition 1, we may now express the estimator as:

$$\hat{t}_G = \sum\nolimits_F y_k I_k / \pi_k E_k\,, \qquad E_k = I\{\Gamma_k = G\}$$

Using the assumption of a sample selection procedure consistent with the sampling design,

$$E(I_k/\pi_k\, E_k |\, D) = \frac{1}{\pi_k} E(I_k \mid D) E(E_k \mid D, I_k = 1) = E(E_k \mid D, I_k = 1)$$

Thus, by independence $E(I_k/\pi_k\, E_k) = E(\, E_k)$, and we conclude that

$$E(\hat{t}_G) = E\left(\sum\nolimits_{\Gamma_k = G} y_k\right) = t_G + E(CovErr_G\,)$$

For the variance estimator, using the above expression for $\hat{t}_G$:

$$V(\hat{t}_G) = \sum\nolimits_F \sum\nolimits_F y_k y_l \big(E\big(\check{I}_k \check{I}_l E_k\, E_l\,\big) - E\big(\check{I}_k E_k\,\big) E(\check{I}_l E_l\,)\big)$$

Also, since $\pi_{k,l} > 0$ for all $y_k y_l \neq 0$

$$E\left(\hat{V}(\hat{t}_G)\right) = \sum\nolimits_F \sum\nolimits_F y_k y_l \big(E\big(\check{I}_k \check{I}_l E_k\, E_l\big) - E(I_k I_l E_k\, E_l/\pi_{k,l})\big)$$

Now, $E\big(\check{I}_k E_k\big) = E(\, E_k)$ and $E(I_k I_l E_k\, E_l/\pi_{k,l}) = E(E_k\, E_l)$ due to independence, and since

$$E(I_k I_l E_k\, E_l/\pi_{k,l} |\, D) = E(E_k \mid D, I_k = I_l = 1)$$

Consequently,

$$E\left(\hat{V}(\hat{t}_G)\right) - V(\hat{t}_G) = -\sum_F \sum_F y_k y_l \big(E(E_k\, E_l) - E(\,E_k)E(\,E_l)\big)$$

verifying that

$$bias\left(\hat{V}(\hat{t}_G)\right) = -V\left(\sum_{\Gamma_k = G} y_k\right) = -V(CovErr_G)$$

Proof of Proposition 4. The proposed estimator can be expressed as:

$$\hat{t} = \frac{N_{frame}}{n} \sum_F y_k\, I\{k \in S\}$$

Since S is a random sample of n units among F, it follows that $P(k \in S) = n/N_{true}$. Thus,

$$E(\hat{t}) = \frac{N_{frame}}{N_{true}} \sum_F y_k = \frac{N_{frame}}{N_{true}}\, t$$

or in other words, $E(\hat{t})/t - 1 = N_{frame}/N_{true} - 1$.

References

Beaumont, J.-F., & Rivest, L.-P. (2009). Ch. 11. Dealing with Outliers in Survey Data. i D. Pfeffermann, & C. R. Rao, *Handbook of Statistics Volume 29A.*

Colledge, M. J. (1995). Frames and business registers: an overview. i B. Cox, D. Binder, A. Chinnappa, A. Christianson, M. Colledge, & P. Kott, *Business Survey Methods* (p. 21-47). New York: John Wiley & Sons, Inc.

Eurostat. (2010). *Business Registers. Recommendations manual.*

Hedlin, D., & Wang, S. (2004). Feeding back information on ineligibility from sample surveys to the frame. *Survey Methodology*, 167-174.

Hidiroglou, M. A., & Lavallée, P. (2009). Ch. 17. Sampling and Estimation in Business Surveys. i D. Pfeffermann, & C. R. Rao, *Handbook of Statistics Volume 29A.*

Kalton, G. (2009). Ch.5. Designs for Surveys over Time. i D. Pfeffermann, & C. R. Rao, *Handbook of Statistics Volume 29A.*

Ohlsson, E. (1992). *SAMU The System for Co-ordination of Samples from the Business Register at Statistics Sweden - A Methodological Description.* Stockholm: Statistics Sweden R&D Report.

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. i B. Cox, D. Binder, A. Chinnappa, A. Christianson, M. Colledge, & P. Kott, *Business Survey Methods.* New York: John Wiley & Sons, Inc.

ONS. (2001). *Review of the Inter-Departmental Business Register.* London: ONS.

Smith, P. (2013). Sampling and Estimation for Business Surveys. i G. Snijkers, G. Haraldsen, J. Jones, & D. K. Willimack, *Designing and Conducting Business Surveys.* Wiley & Sons.

Srinath, K. (1987). Methodological problems in designing continuous business surveys: some Canadian experiences. *Journal of Official Statistics*, 283-288.

Srinath, K. P., & Carpenter, R. M. (1995). Sampling methods for repeated business surveys. i B. Cox, D. Binder, A. Chinnappa, A. Christianson, M. Colledge, & P. Kott, *Business Survey Methods* (p. 171-183). New York: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag.