RETAINING ELECTRONIC REPORTERS IN BUSINESS SURVEYS: ESTIMATION IMPACT

Chantal Grondin, M.A. Hidiroglou, and Pierre Lavallée
M.A. Hidiroglou, 11-A, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

1. INTRODUCTION

The Survey of Employment, Payrolls, and Hours (SEPH) is designed to provide monthly estimates to measure levels and month-to-month trends of payroll employment, paid hours and earnings. The data are compiled at detailed industrial levels for Canada, provinces and territories. The target population is composed of all employers in Canada, except those primarily involved in agriculture, fishing and trapping, private household services, religious organisations and military personnel of defence services. In all, there are close to one million establishments in scope for the survey.

Considerable savings are being realised with this survey since employment and gross monthly payrolls variables are obtained directly each month from the payroll deductions remittance (PD7) forms from the Canada Customs and Revenue Agency (CCRA). To complement this administrative source of data, an independent sample of approximately 11,500 establishments is used every month to collect information on employment, gross monthly payrolls, total hours, summarised earnings, as well as the allocation of these variables by categories of employees (paid by the hour, salaried and other). This complement is known as the Business Payroll Survey (BPS). The frame used to draw the sample is the Business Register, which is an electronic database of all Canadian businesses maintained at Statistics Canada. The sample is maintained for births and deaths, and rotation of one twelfth of the sample occurs every month. The sample design is stratified by size (based on employment), geography (by region), and industry groups (sub-sectors of the North American Industrial Classification System, or NAICS 3-digit).

The SEPH final estimates use a combination of both data sources. Regression models are used to predict hours and summarised earnings from the sample of respondents and estimated regression coefficients are applied to the administrative variables to mass impute hours and summarised earnings for every administrative unit. All other variables are derived by multiplying employment, hours or summarised earnings by a ratio (or function of ratios) estimated from the BPS sample. More details about the methodology of SEPH are available in Rancourt and Hidiroglou (1998), and Grondin (2000).

The quality of monthly estimates produced by SEPH depends largely on the quality of the regression coefficients and ratios estimated from the sample of establishments. The quality of these coefficients could be improved by increasing the current sample size. Unfortunately, the cost of data collection and the resulting increase in response burden are prohibitive. However, some businesses are, more and more, willing to respond to the survey on an electronic basis. These businesses are known as EDRs (or *Electronic Data Reporters*). This fairly new collection method has the advantage of being inexpensive and the associated response burden is minimal. Hence, increasing the number of EDRs in the sample sounds promising in terms of increasing total sample size without increasing collection costs. On the other hand, the problem with increasing the sample size this way is that since the EDR status of businesses is only available from sampled units, there is the risk of introducing bias into the estimates.

In this paper, we propose a number of ways to increase the current establishment sample size with EDR units, and determine the associated bias.

2. TWO SAMPLING OPTIONS FOR EDRs

SEPH uses a stratified design with rotation. However we confine ourselves to a simplified version of the design to ease the development. We assume simple random sampling without replacement from a population that is static, meaning that we pretend that there are no deaths, births or changes in structures of the sampling units, and no changes in the stratification variables. Furthermore, we assume that there is no non-response.

We represent the population for a given cycle *c* of the survey as *U* (of size *N*) and the corresponding sample as s_c. We confine our discussion to two cycles: start-up ($c=1$) and the next occasion ($c=2$). At time of start-up let the sample size of s_1 be *n*. We observe that there are *t* EDR units in this sample and *n-t* non-EDR (regular) units. Several options are possible with respect to the treatment of the EDRs and their associated weights for sample s_2 (the sample at cycle2).

We proceed to describe two options (options A and B) for treating the EDRs after they have been identified. The main difference in the treatment of the EDRs between these two options is how the sampling procedures are affected and how the sampled units are weighted.

2.1 Option A

The *t* EDRs discovered during cycle 1 are retained in the sample at cycle 2, and *t* non sampled population units are selected for cycle 2 to increase the net number of units to *n*+*t*.

Assuming that the population is static, the sampling weights for cycle 2 are $N/(n+t)$ for each unit. Some of the *advantages* of this procedure include the following: (i) the process appears transparent because all you see are sampling fractions gradually increasing with time; (ii) the weights are relatively stable over time since they only vary when the sample is increased on account of having found more EDRs. However, some of the *disadvantages* are: (i) retaining EDRs in the sample and increasing sampling fractions become quite unwieldy because of the implemented sample rotation. Rotation schemes operate using a continuous sampling window of length $f=n/N$ that moves across the (0,1) interval. Retaining EDRs in the sample can then become an operational problem, as the sampling window will eventually consist of the union of a closed interval and several points (the EDRs) on the (0,1) interval. It should be noted that rotating units into the sample must exclude any existing EDR. As well, sampling fractions must take into account the number of EDRs already in the sampling window, making the whole process an iterative one; (ii) the scheme is biased, as shown in section 3, and this bias will slowly increase over time as we add EDRs. The size of this bias is a function of the difference between the characteristics of the EDR and the non-EDR units, as well as of the relative proportion of EDRs in the population.

2.2 Option B

The *t* EDRs discovered during cycle 1 are moved to a take-all stratum at the beginning of cycle 2, forcing them to remain in the sample with a weight of one. Then, *t* non-sampled population units are added to the sample yielding a net number of units for cycle 2 of *n*+*t*.

This scheme has the following *advantages*: (i) management of the sampling fractions, even with sample rotation, is much simpler than with the other option; (ii) sampling fractions need to be computed only once a month (unlike the iterative process needed with option A); (iii) if rotation is part of the sampling scheme, the bias increases as we add EDRs to the sample, but then drops and tends towards zero as rotation cycles through the population.

Disadvantages of this scheme are: (i) initially, there could be some instability in the sampling weights as the weight of an EDR is reduced to one; (ii) assigning a weight of one to EDR units will have a small impact on the estimates and this contradicts the idea of adding more units to the sample to improve estimates.

We proceed to investigate the biases associated with these two options in section 3.

3. BIASES ASSOCIATED TO EACH OPTION

We evaluate the bias of each option for two estimation schemes: totals, and ratios. Let $\boldsymbol{y}=(y_1,\ldots y_p)$ denote the *p*-dimensional vector of observations collected by the sample. We confine ourselves to $p=2$, to illustrate the bias associated to the ratio estimator.

For a given occasion, let $\hat{\boldsymbol{Y}}=\sum_s w_k \boldsymbol{y}_k$ be the estimated total for the multivariate characteristic $\boldsymbol{y}$, and w_k be the weight associated with each unit *k* in the sample *s* at a given survey cycle. If estimating ratios is also of interest, then the estimated ratio is $\hat{R}=\hat{Y}_1/\hat{Y}_2$. The relative bias of the ratio estimator will be much smaller that the relative bias of the estimated totals defining it. Denote the bias associated with each estimated total $\hat{Y}_p$ as $B(\hat{Y}_p)$, p=1,2. Then the bias of $\hat{R}$, using the first order Taylor expansion is $B(\hat{R})/R \doteq \left(B(\hat{Y}_1)/Y_1 - B(\hat{Y}_2)/Y_2\right)$ which is much smaller than $B(\hat{Y}_p)/Y_p$. Generally speaking, this implies that the bias of a smooth function of totals (i.e. regression estimation) can be connected (via a Taylor expansion) to the biases of its individual components.

For a given *y*-variable of interest, we decompose the corresponding population total $Y=\sum_{i=1}^{N} y_i$ into two parts, denoted as Y_E and Y_R. Here Y_E represents the total of the *y* variable for the T ($T>1$) EDR units in the population, while Y_R represents the total of the *y* variable for the regular units (non-EDR) in the population. We assume that $n>T$. Suppose that $t \geq 1$ EDRs are found in the sample s_1 at cycle 1. The sample s_1 can be split into a part s_{1,E_t} that contains *t* EDRs, while the other, $s_{1,R_{n-t}}$, contains *n-t* regular units. At cycle 2, the sample s_2 is similarly

split into s_{1,E_t} with t EDR units and s_{1,R_n} with n regular units. Let the set of all possible $\binom{N}{n}$ samples s_1 drawn at cycle 1 be denoted as ς_1. Similarly, let the set of all possible samples s_2 at cycle 2 resulting from the action on the EDR units be denoted as ς_2. It should be noted that ς_2 will be different from ς_1, since a realised sample s_2 at cycle 2 may be different from the corresponding first sample s_1 at cycle 1 if s_1 contains EDR(s).

3.1 Option A

For this option, the weight changes from *N/n* at cycle 1 to *N/(n+t)* at cycle 2 if $t \geq 1$. The resulting estimator at cycle 2 is:

$$\hat{Y}^*_{s_2} = \frac{N}{n+t}\sum_{i=1}^{n+t} y_i \ .$$

We have that the conditional expectation of $\hat{Y}^*_s$ for $t \geq 0$ is given by:

$$\begin{aligned} E\left(\hat{Y}^*_{s_2} \mid t\right) &= E\left[\frac{N}{n+t}\sum_{i=1}^{n+t} y_i \mid t\right] \\ &= \frac{N}{n+t}E\left[\left(\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{n+t} y_i\right)\mid t\right] \\ &= \frac{Nn}{n+t}\overline{Y}_R + \frac{Nt}{n+t}\overline{Y}_E \end{aligned}$$

Here $\overline{Y}_R$ and $\overline{Y}_E$ respectively denote the population means of the EDR units and non-EDR units.
The unconditional expectation of $\hat{Y}^*_{s_2}$ requires taking the expectation of the linearised (via Taylor) of version of the random variables $t/(n+t)$ and $1/(n+t)$ (via Taylor). We have that

$$E\left(\hat{Y}^*_{s_2} \mid t\right) \cong N\left(1-\frac{t}{n}+\frac{t^2}{n^2}\right)\overline{Y}_R + \frac{N\,t}{n}\left(1-\frac{t}{n}\right)\overline{Y}_E$$

where t / *n* is distributed as a hyper-geometric distribution. Its mean and variance are respectively *T/N* and $\frac{T(N-T)}{n\,N^2}\ \frac{N-n}{N-1}$. Using these results, the approximate unconditional expectation of $\hat{Y}^*_{s_2}$ can be shown to be

$$E\left(\hat{Y}^*_{s_2}\right) \doteq Y + \frac{N}{N-T}\left(\overline{Y}-\overline{Y}_E\right)\frac{T}{n(N-1)}\left[(N-T)-n+nT\right]$$

Hence, the approximate bias of $\hat{Y}^*_{s_2}$ is:

$$\begin{aligned} Bias\left(\hat{Y}^*_{s_2}\right) &= \frac{N}{N-T}\left(\overline{Y}-\overline{Y}_E\right)\frac{T}{n(N-1)}\left[(N-T)-n+nT\right] \\ &\doteq \frac{NT}{N-T}\,\frac{1}{n}\left[(1-f)+fT\right]\left(\overline{Y}-\overline{Y}_E\right) \end{aligned}$$

3.2 Option B

We first suppose that there is only one EDR in the population. The bias of the associated estimator is provided in sub-section 3.2.1. In general, however, there could be $T\ (T>1)$ EDRs in the population, and the bias of the resulting estimator is given in sub-section 3.2.2.

3.2.1 The case of one EDR in the population (T=1)

We first suppose that there is only one EDR (say *k*) in the population. Two configurations are possible: the EDR is either included in the sample s_1 (config.1) or it is not included (config.2). Figure 1 represents this.

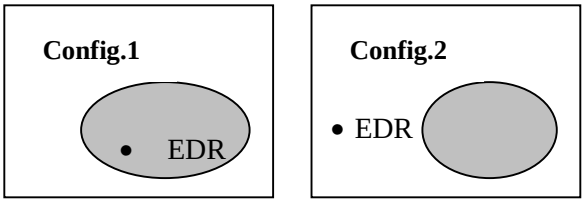

Figure 1: The case of a single EDR

In configuration 1, the EDR unit is included in the sample s_1 at cycle 1. An extra unit is selected from the non-sampled portion of the population. The resulting number of regular units at the start of cycle 2 will still be *n*. The EDR unit is assigned a weight of one at cycle 2 while the weight of the non-EDR units are adjusted so that the sum of the sample weights for all the units in s_2 adds up to *N*.

In configuration 2, there is no increase in the sample size because the EDR unit is not identified. The number of regular units remains at *n* for cycle 2. The sampling weight of the non-EDR units remains at *N*/*n*.

The population total $Y = \sum_{i=1}^{N} y_i$ is estimated at cycle 2 by $\hat{Y}^*_{s_2}$, where:

$$\hat{Y}^*_{s_2} = \begin{cases} y_k + \dfrac{N-1}{n}\sum_{s_{2,R_n}} y_i & \textit{for config 1} \\ \dfrac{N}{n}\sum_{s_2} y_i & \textit{for config 2} \end{cases} \qquad (3.1)$$

The probabilities of selection of s_2 depend on whether the EDR unit was included or excluded in s_1.

Configuration 1: If the EDR unit was included in s_1, then an additional unit is selected from the $(N-n)$ non-sampled population units: there are $\binom{N-1}{n-1}$ such samples. The resulting number of district s_2 samples is $\binom{N-1}{n}$. This is equivalent to drawing *n* units (excluding the EDR) from a population of size– N-1 (original population excluding the EDR). Hence the probability of drawing s_2 is equal to the product of the probability drawing samples s_1 that contain the EDR times the probability of drawing the resulting s_2 samples. That is

$$p(s_2) = p(s_1 \ni k)\, p(s_2 \mid s_1 \ni k) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \frac{1}{\binom{N-1}{n}}$$

Configuration 2: If the EDR unit was not included in s_1, then $s_2 = s_1$ and $p(s_2) = 1/\binom{N}{n}$: there are $\binom{N-1}{n}$ such samples.

The probabilities of selection for the samples s_2 are

$$p(s_2) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \frac{1}{\binom{N\text{-}1}{n}} \quad \textit{if} \quad k \in \varsigma_{2,E_1}, \text{ and}$$

$$p(s_2) = \frac{1}{\binom{N}{n}} \quad \textit{if } k \in \varsigma_{2,E_0} \text{ where } \varsigma_{2,E_1} \text{ and } \varsigma_{2,E_0} \text{ are}$$

respectively the subsets of ς_2 that either contain or do not contain the EDR. The expectation of (3.1) is:

$$E\left(\hat{Y}^*_{s_2}\right) = \sum_{s_2 \in \varsigma_2} p(s_2)\ \hat{Y}^*_{s_2} = \sum_{s_2 \in \varsigma_{2,,E_1}} p(s_2)\, \hat{Y}^*_{s_2} + \sum_{s_2 \in \varsigma_{2,,E_0}} p(s_2)\ \hat{Y}^*_{s_2} \qquad (3.2)$$

leading to

$$E\left(\hat{Y}^*_{s_2}\right) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \frac{1}{\binom{N-1}{n}} \sum_{s_2 \in \varsigma_{2,E_1}} \left(y_k + \frac{N-1}{n} \sum_{s_2(\ni k)} y_i \right) + \frac{1}{\binom{N}{n}} \sum_{s_2 \in \varsigma_{2,E_0}} \left(\frac{N}{n} \sum_{s_2(\ni k)} y_i \right)$$

To evaluate the above sums, we first work out in how many samples y_k will appear. There are $\binom{N-1}{n}$ such samples, because the value y_k is only present in samples s_2 that belong to ς_{2,E_1}. Second, given that y_k is present in s_2, $y_j\,(j \neq k)$ will appear in $\binom{N-2}{n-1}$ samples. Finally, if y_k was not present in s_2, then the value $y_j\,(j \neq k)$ can only appear in $\binom{N-2}{n-1}$ samples. Hence,

$$E\left(\hat{Y}^*_{s_2}\right) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} \frac{1}{\binom{N-1}{n}} \left[\binom{N-1}{n} y_k + \binom{N-2}{n-1} \frac{N-1}{n} \sum_{U(\ni k)} y_i \right] + \frac{\binom{N-2}{n-1}}{\binom{N}{n}} \frac{N}{n} \sum_{U(\ni k)} y_i$$

$$= \frac{n}{N} y_k + \frac{N-1}{n-1} \frac{n}{N} \frac{n-1}{N-1} (Y - y_k)$$

That is, $E\left(\hat{Y}^*_{s_2}\right) = \dfrac{(N-f_1)}{(N-1)} Y - \dfrac{N(1-f_1)}{(N-1)} y_k$, where $f_1 = n/N$ is the original sampling fraction in sample s_1. The bias of $\hat{Y}^*_{s_2}$ is therefore $\dfrac{N(1-f_1)}{(N-1)}\left(\bar{Y} - y_k\right)$. Note that the bias is negligible if the EDR value y_k is close to the overall sample population mean $\bar{Y}$.

3.2.2 The case of several EDRs in the population (T>1)

Let ς_{1,E_t} denote the set of all possible samples s_1 that contain $t \geq 0$ EDR units in cycle 1: ς_{1,E_0}. Similarly, ς_{2,E_t} $(t = 0,1,...,T)$ denotes the set of all possible samples that contain $t \geq 0$ EDR units in cycle 2. Recall that if we have discovered *t* EDR units in cycle *1*, then the sample size is increased by *t* units between cycles 1 and 2, yielding the sample s_2 at cycle 2.

We estimate the population total *Y* at cycle 2 as:

$$\hat{Y}^*_{s_2} = \begin{cases} \sum_{s_{2,E_t}} y_k + \dfrac{N-t}{n} \sum_{s_{2,R_n}} y_k & \textit{for config 1} \\ \dfrac{N}{n} \sum_{s_{2,R_n}} y_k & \textit{for config 2} \end{cases}$$

where s_{2,E_t} and s_{2,R_n} denote sample realisations with *t* sampled EDR units (*t=0,1,…,T*).

Once more $E\left(\hat{Y}^*_{s_2}\right) = \sum_{s_2 \in \varsigma_2} p(s_2)\ \hat{Y}^*_{s_2}$, and this can be decomposed into EDR and non-EDR portions as:

$E\left(\hat{Y}^*_{s_2}\right) = \sum_{t=0}^{T} \sum_{s_{2,t} \in \varsigma_{2,E_t}} p(s_2)\hat{Y}^*_{s_2}$ where ς_{2,E_t} $(t = 0, 1, ..., T)$ denotes the set of all possible samples s_2 that contain $t \geq 0$ EDRs ,and $s_{2,t}$ are the s_2 samples that contain *t* EDRs. Note that ς_2 is simply the union of the ς_{2,E_t}'s. The probability that a sample $s_{2,t}$ will contain *t* EDRs is

$$\begin{aligned} p(s_{2,t}) &= p(s_1 \ni t\ EDRs)\ p(s_{2,t} \mid s_1 \ni t\ EDRs) \\ &= \frac{\binom{N-T}{n-t}\binom{T}{t}}{\binom{N}{n}} \frac{1}{\binom{N-n}{t}} \end{aligned}$$

Hence

$$E\left(\hat{Y}^*_{s_2}\right) = \sum_{t=0}^{T} \frac{\binom{N-T}{n-t}\binom{T}{t}}{\binom{N}{n}\binom{N-n}{t}} \sum_{s_{2,t} \in \varsigma_{2,E_t}} \left(\sum_{s_{2,E_t}} y_k + \frac{N-t}{n} \sum_{s_{2,R_n}} y_k \right)$$

The above expression can be simplified to:

$$E\left(\hat{Y}^*_{s_2}\right) = \frac{(N-Tf)}{(N-T)} Y - \frac{(N-n)}{(N-T)} Y_E$$

Thus, the bias introduced by retaining *t* EDR units (amongst the *T* EDRs of the population) into the sample is:

$$\text{Bias}(\hat{Y}_{s_2}) = \frac{\text{NT}(1-\text{f})}{(\text{N}-\text{T})}\left(\overline{Y} - \overline{Y}_E\right)$$

The bias tends towards 0 when the EDR mean ($\overline{\boldsymbol{Y}}_{\boldsymbol{E}}$) approaches the overall population mean ($\overline{\boldsymbol{Y}}$), or if the sampling fraction tends towards 1. The magnitude of this bias depends on knowledge of *T*, and the difference $\overline{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}_{\boldsymbol{E}}$. In practice, we do not know exactly the number of EDRs in the population (*T*). However, *T* can be "guestimated" after several cycles of a survey, or it can be bounded above and below (as Rao 1985 suggests), thereby providing an idea of the span of the bias.

3.2.3. Arbitrary Weights for the EDRs

In general, we could assign a weight of $1 \leq w_E \leq \dfrac{N}{n}$ to the EDR units and $w_A = \dfrac{N - t\ w_E}{n - t}$ to the non-EDR units: note that these weights sum up to the population size *N*. The estimator is then:

$$\hat{Y}^*_{s_2} = \begin{cases} w_E \sum_{s_{2,E_t}} y_k + w_A \sum_{s_{2,R_n}} y_k & \textit{for config 1} \\ \dfrac{N}{n} \sum_{s_{2,R_n}} y_k & \textit{for config 2} \end{cases}$$

We use conditional expectations to obtain the unconditional expectation of $\hat{Y}^*_{s_2}$ as it greatly simplifies the algebra. We can write $\hat{Y}^*_{s_2}$ as $w_E + \bar{y}_E + w_A (N-t)\bar{y}_A$

Hence

$$E\left(\hat{Y}^*_{s_2} \mid t\right) = t\, w_E\, \overline{Y}_E + w_A\, (N-t)\overline{Y}_A$$

and

$$\begin{aligned} E\left(\hat{Y}^*_{s_2}\right) &= E\left[E\left(\hat{Y}^*_{s_2} \mid t\right)\right] \\ &= \frac{nT}{N} w_E\ \overline{Y}_E + \left(N - \frac{nT}{N}\right) w_A\ \overline{Y}_A \end{aligned}$$

which simplifies to

$$E\left(\hat{Y}_s^*\right) = \frac{\left(N - T\ f\ w_E\right)}{\left(N - T\right)} Y - \frac{\left(N - n\, w_E\right)}{\left(N - T\right)} Y_E$$

The resulting bias is:

$$Bias(\hat{Y}_s) = \frac{NT(1 - f\ w_E)}{(N - T)}\left(\overline{Y} - \overline{Y}_E\right).$$

The significance of the above result is that if cost comes into play and EDR units do cost to collect (say a portion α of the non-EDR units), then their existence in the sample implies that they cannot be replaced on a one to one basis with the non-EDR units. Rather, to keep overall sample costs fixed, the sample that consists of non-EDR units at cycle 1 can be increased with *n t* (1- α) units.

4. APPLICATION TO THE SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS

For our specific application (SEPH), an attempt was made to identify potential EDR units in the BPS sample. This was done through the addition of an extra question in the June 1999 survey questionnaire, asking which payroll service or system produces the company payroll. The aim of the question was to identify establishments whose payroll service made them potential candidates for conversion to the EDR collection mode. We will call these establishments "potential units". Some 353 establishments out of 10,450 reported using a system easily convertible to EDR. This represents only 3.4 % of the sample (non-weighted). Yet, since not all these potential units will agree to be converted to the EDR mode, we know that the percentage of real EDRs in the sample will be less than 3.4%. However, since the BPS units in take-some strata undergo a one-twelfth rotation each month, it means that approximately 960 new units rotate into the sample each month. Thus, in the course of a year, about 390 new EDRs could potentially be identified.

A study was carried out to compare the characteristics of potential EDRs versus regular units, based on the information collected for the month of June 1999 We first examined the effect of the potential units on regressions models used for constructing the SEPH estimates. It was found that they have no significant impact on regression coefficients. However, it was noted that the potential units had significantly different numbers of employees as compared to regular establishments. This difference was not large enough to warrant re-stratifying the population into EDRs and non-EDRs.

Furthermore, analysis showed that the potential EDR units were significantly different from regular units, for hours worked and number of part-time employees.

5. CONCLUSION

The conversion of units contacted via a regular mode of data collection such as mail or telephone to one of Electronic Data Reporting (EDR) implies cost savings. These cost savings can be re-invested into the sample by increasing its size. However, this also implies potential biases. Two options were proposed. In one option (option A), the weights of all units are re-computed to reflect the increased sample size: the bias of the estimator of total is of order 1/*n*. For option B the weights of the EDR units are set equal to one, and the weights of the remaining in-sample units are adjusted: the bias of the corresponding estimator of total is of order one. The degree of bias mainly depends on the differences between the means of the ERR and of the regular units.

The choice between those two options depends on how much potential bias one is willing to accept versus how difficult it is to implement the chosen procedure. In the case of a survey that has rotation, the bias will decrease with both options.

REFERENCES

Grondin, C. (2000). "Using Electronic Data Reporters in the BPS", *Internal working paper*, May 23, 2000.

Hidiroglou M.A. and Srinath, K.P. (1981). "Some Estimators of Population Totals from Simple Random Samples Containing Large Units". *Journal of the American Statistical Association*, Vol. 76, No. 375, 690-695.

Rancourt, E. and Hidiroglou M. (1998). Use of Administrative Records in the Canadian Survey of Employment, Payrolls, and Hours, *Proceedings of The Survey Section of the Canadian Statistical Society*, 39-49.

Rao, J.N.K. (1985). Conditional Inferences in Survey Sampling. *Survey Methodology*, 15-32.