Managing Response Burden by Controlling Sample Selection and Survey Coverage

Sébastien Landry[1]
[1]Statistics Canada, Business Surveys Methods Division, R.H. Coats Building, 11th floor, Ottawa, ON, Canada, K1A 0T6

Abstract
Statistical agencies are constantly making efforts to control the response burden of their household and business survey respondents. Statistics Canada's Survey on Employment, Payroll and Hours is no exception. This monthly business survey, which produces estimates and determines the month-to-month changes for variables such as employment, earnings and hours at detailed industrial levels for Canada, the provinces and the territories, currently manages response burden by making use of administrative data and by having rules that prevent establishments from rotating in the sample too soon after being rotated out. Recently, two new ideas to decrease even more the response burden for respondents to this survey have been studied. The first is to control the overlap of the samples from one month to the next by the use of the microstrata method (Rivière (2001)) in the sample selection process. The second is the increased number of establishments in the take-none strata. This paper will present the studies that evaluated the pros and cons of implementing each of these new features in the survey.

Key Words: Business surveys, Response burden, Sample selection, Take-none strata

1. Introduction

With the ever increasing demand for data being placed upon surveys, it has become essential that statistical agencies manage the response burden of their survey respondents in order to increase the likelihood they will participate in surveys. Successful response burden management can be achieved by using administrative data when possible, designing efficient questionnaires and developing sampling designs to control response burden. Statistics Canada's Survey on Employment, Payroll and Hours (SEPH) has already implemented methods to manage response burden. However, recent proposals to improve the SEPH's response burden management through its sampling design have been proposed. The purpose of this paper is to present these proposals and the studies that have been made to evaluate their potential benefits.

The SEPH methodology and its former response burden management strategy (before the proposals) will be presented in section 2. The first proposal to improve response burden management, which deals with sample selection control, will be shown in section 3. The second proposal to improve response burden management, which relates to the size of the take-none strata, will be explained in section 4. Section 5 will summarize and conclude this paper.

2. SEPH Methodology

The main goal of the SEPH is to provide data on employment, earnings and the number of hours worked. Estimates are required at the industrial and provincial/territorial levels. To obtain the desired results, the SEPH relies on a survey called the Business Payroll Survey (BPS), which will be introduced in section 2.1, and on administrative data, which is part of the SEPH's response burden management strategy that will be explained in section 2.2.

2.1 BPS Methodology

The BPS is a monthly establishment survey that selects its sample from Statistics Canada's Business Register using a stratified sampling design. Its primary objective is to send a sample of 15,000 establishments into collection. The selected establishments stay in the sample for 12 consecutive months and, each month, 1/12 of the sample is replaced through a rotation scheme. The variables collected by the BPS include aggregates, such as number of salaried employees and number of employees paid by the hour, and ratios, such as average weekly earnings and average weekly hours. One important requirement of the BPS methodology is that the enterprise (the highest level of Statistics Canada's statistical business structure), which can have many establishments, is the survey respondent and has to provide the required information for all its selected establishments.

For a more detailed description of the SEPH methodology, please read Morin (2010).

2.2 SEPH Response Burden Management Strategy before the Proposals

The former SEPH response burden management strategy included three components: the use of administrative data, sampling selection rules and take-none strata.

2.2.1 Administrative Data

The Canada Revenue Agency and Statistics Canada's Public Sector Statistics Division provide the SEPH program with files that cover the whole survey frame. The variables selected from these files for an establishment are its number of employees and the monthly earnings of its employees. These files are then used to produce domain totals and to calibrate the weights according to the auxiliary variables previously defined.

2.2.2 Sampling Selection Rules

Two selection rules, named and explained below, are enforced to manage the response burden of the enterprises which have establishments selected in the sample.

- Exclusion rule: once establishments from an enterprise are selected in the sample, no other non-selected establishment from the enterprise can be selected in the sample. They are deemed excluded.
- Freeze rule: once the selected establishments from an enterprise are rotated out of the sample, no establishment from the enterprise can be selected in the sample for the following 12 months. They are deemed frozen.

In addition to these selection rules, a size measure, defined below, has been created to allow increasing the proportion of births in the sample and decreasing the number of selected establishments that had left the sample fairly recently. Therefore, the selection of the portion of the sample being rotated in was effectively done using a design with probability proportional to the size measure.

$$\text{Size} = \begin{cases} 12 & \text{if the establishment is a birth} \\ 0 & \text{if the establishment is excluded or frozen} \\ x/24 & \text{if the establishment left the sample } 12+x \text{ months ago, } 0<x<24 \\ 1 & \text{otherwise.} \end{cases}$$

2.2.3 Take-None Strata

It is common practice for business surveys to systematically exclude the smallest establishments from the sample since they do not contribute much to estimates of totals. The SEPH being mostly interested in estimating ratios, it was felt that small establishments did contribute to the estimates. Nevertheless, take-none strata have been introduced in 2008 to avoid sampling small establishments that were often unusable at the estimation stage mainly because they were flagged as out-of-scope during collection. As a result, establishments with 0 or 1 employee were sent to take-none strata.

Although establishments in take-none strata cannot be sampled, they are taken in account at the estimation stage since they are included in the calculation of auxiliary variable totals that are used in calibration.

3. Improving Response Burden through Sample Selection Control

The sampling design used probability proportional to size sampling for the rotated-in portion of the sample, as described in section 2.2.2. This strategy was somewhat successful in controlling the response burden of establishments being recently rotated out of the sample. However, with this approach, establishments can still be resampled right after the freeze period. As a result, some establishments did get selected in the sample not long after being rotated out. Some of these respondents complained about being reselected in the sample. The proposal to address this issue was to implement the microstrata method as explained by Rivière (2001).

3.1 Microstrata: the Method

The microstrata method was created to coordinate the selection of a sample according to the cumulative burden from previous surveys or occurrences of a survey of the establishments. Usually, a burden value is added to the cumulative burden of an establishment when it is selected in the sample. The method is based on permanent random numbers (PRN) going from 0 to 1 and, basically, the strategy consists of always selecting the lowest PRN in the sample. The PRN remain the same (except when births or deaths occur), but the establishment assigned to a PRN can change for every survey or occurrence, depending on the desired kind of sampling coordination. In the case of negative coordination, establishments within a microstratum are sorted according to their cumulative response burden and the larger PRN are given to establishments with larger cumulative burden. The microstrata are defined by the intersection of the strata of the previous surveys on which sampling coordination is being done.

It was felt that the microstrata method could be beneficial for the SEPH because it would make the period between two selections of an establishment more uniform within a stratum. Also, the method keeps the original sampling design properties.

3.2 SEPH Implementation of the Microstrata Method

The way the SEPH intends to implement the microstrata method mimics in a way the positive coordination with rotation rate scheme from Rivière (2001). So, a burden value

of 1 would be given to establishments when they are rotated out of the sample; the other establishments would receive a burden value of 0. Furthermore, the microstrata for the current month survey would be formed by the strata from the previous month survey. Births from one stratum would be assigned new PRN in a way that they would be uniformly distributed within the [0,1) interval. Also, the exclusion and freeze rules would still apply, which means that an excluded or frozen establishment would still not be selected in the sample even if it had the lowest PRN available.

Figures 1 and 2 provide a simple example of how the microstrata method would work on one SEPH stratum that keeps the same establishments from the previous month to the current month. The stratum is represented by a [0,1) interval with the establishments standing on their assigned PRN along the interval.

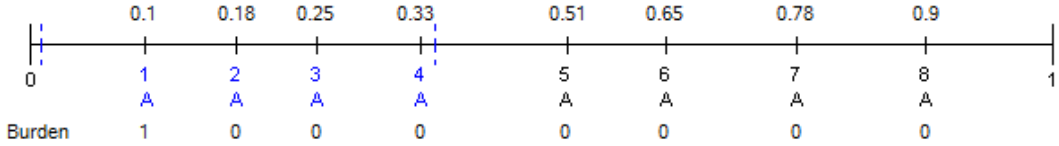

Figure 1: Microstrata example: stratum with previous month PRN and current month burden

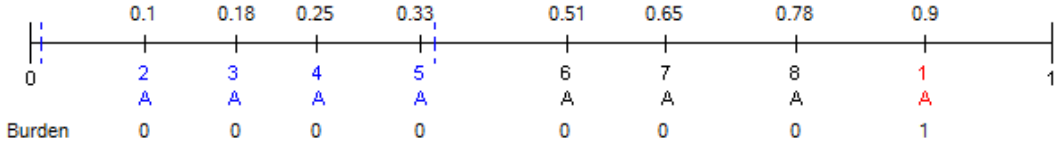

Figure 2: Microstrata example: stratum with current month burden and PRN

In this example, establishments 1A, 2A, 3A and 4A (in blue) were in the sample for the previous month and establishment 1A is being rotated out for the current month while establishments 2A, 3A and 4A remain in the sample. The burden values assigned to the establishments in figure 1 reflect this. Since the microstratum is equivalent to the stratum in this case, the establishments are sorted according to their cumulative burden. Establishment 1A (red denotes a frozen establishment), being the establishment with the highest cumulative burden, is assigned the highest PRN, as seen in figure 2. The other establishments are assigned the lower PRN next to their previous PRN. If an additional establishment needs to be selected in the sample, it will be 5A, which is the one with the lowest PRN not already in the sample and available for selection.

3.3 Simulation Study

In order to verify if the potential benefits of implementing the microstrata method would materialize, the microstrata method has been simulated over the SEPH sampling process for the months from August 2006 to September 2009. The criterion used to determine whether the microstrata method or the sampling selection process before the proposals managed better the response burden was the number of establishments that have been rotated in the sample more than once by the sampling selection scheme during the simulation period. Due to the freeze rule, only establishments rotated in the sample between August 2006 and September 2007, around 12,000 establishments, were eligible to be rotated in more than once during the simulation period. The microstrata method, with 31 establishments selected more than once during the simulation period, did better than the sampling selection process before the proposals which produced 47 establishments selected more than once during the simulation period. The difference

between the two sampling selection schemes is not very large, but it probably would have been larger if the simulation period would have been longer.

4. Improving Response Burden by Increasing the Size of the Take-None Strata

The current take-none strata have been implemented in the SEPH more for collection issues than response burden management. It was suggested that the take-none strata should be defined from the point of view of response burden management. Since the resulting take-none strata would be larger than the current ones, the issues leading to the implementation of the current take-none strata would still be addressed by the suggested take-none strata.

4.1 Proposal for Take-None Strategy

According to the proposed take-none strategy, the take-none strata would be defined independently for each domain. The take-none boundary (establishments having equal or less employees than the boundary would be sent to take-none strata) for a domain would be the largest number between 1, 2, 3, 4, 5 and 10 for which the employment proportion falling in a take-none stratum would remain under 5% of the domain's total employment. For cases where the employment of establishments having one employee represents more than 5% of the domain's total employment, the take-none boundary would be 1.

The proposed take-none strategy has the obvious benefit of eliminating the response burden for the smallest establishments that would end up in the take-none strata. Also, it could also lead to a decrease in collection costs as the required sample size necessary to obtain estimates with equivalent quality indicators as the current take-none strategy (later referred to simply as sample size) may be smaller.

4.2 Simulation Study

Again, a simulation study was done to measure the effects of the proposed take-none strategy on the sample size, the number of critical strata (this term will be defined in section 4.2.2), the potential bias and the quality of the estimates. In this study, the proposed take-none strategy has been applied to the survey frames from August 2009 to July 2010. Then, the sample size has been recalculated and the establishments from the original sample that would fall into a take-none stratum according to the proposed take-none strategy have been removed from the sample. The resulting estimates simulated the proposed strategy. Note that by doing this, the resulting sample might not match the recalculated sample size.

4.2.1 Sample Size

Table 1 indicates the proportion of the frame falling in the take-none strata, both in terms of establishments and employment, and the sample size for the current and the proposed take-none strategies. As expected, the proposed take-none strategy increased the proportion of the frame falling in the take-none strata. It was also successful at decreasing the sample size by up to 5%.

Table 1: Proportion of Frame in Take-None Strata and Sample Size

Take-none strategy	*Proportion of frame in take-none strata (%)*		*Sample size*
	Establishments	*Employment*	
Current	34.6	2.8	15,000
Proposed	40.2	4.5	14,340

4.2.2 Critical Strata

A stratum is deemed critical if all its establishments are represented by only a few enterprises. Because of the sampling selection rules stated in section 2.2.2, there is a real possibility for these strata that there is no establishment available for selection for a given month because all of them are either excluded or frozen, which means that the required sample size could not be met.

The proposed take-none strategy would create 255 critical strata, which is more than the 214 critical strata created by the current take-none strategy.

4.2.3 Relative Difference in Estimates

Tables 2 and 3 summarize the results of the relative differences of all 3-digit North American Industrial Classification System (NAICS3) domains, both at the Canada and provincial levels, when comparing the take-none strategies. The variables of interest used for this comparison are the average weekly earnings (A_AWE), the number of employees paid by the hour (H_EMPL), the number of salaried employees (S_EMPL) and the average weekly hours worked (A_AWH).

The distribution of the relative differences does not show that the proposed take-none strategy would introduce bias to the estimates.

Table 2: Distribution of the Relative Differences (%) of the Estimates Using the Proposed Take-None Strategy at the Canada Level

Variable	*P95*	*Q3*	*Median*	*Q1*	*P5*
A_AWE	2.44	0.39	0.00	-0.14	-2.24
H_EMPL	10,45	1,10	0,00	-0,18	-7,69
S_EMPL	10,19	1,07	0,00	-1,14	-10,05
A_AWH	1,65	0,24	0,00	-0,13	-1,69

Table 3: Distribution of the Relative Differences (%) of the Estimates Using the Proposed Take-None Strategy at the Provincial Level

Variable	*P95*	*Q3*	*Median*	*Q1*	*P5*
A_AWE	4,14	0,50	0,00	-0,14	-2,95
H_EMPL	13,56	1,18	0,00	-0,28	-9,67
S_EMPL	15,79	1,47	0,00	-1,15	-17,62
A_AWH	2,86	0,33	0,00	-0,16	-2,34

4.2.4 Absolute Relative Difference in Estimates

Tables 4 and 5 summarize the results of the absolute relative differences of all NAICS3 domains, both at the Canada and provincial levels, when comparing the take-none strategies, for the same variables of interest as in section 4.2.3.

The distribution of the absolute relative differences indicates that the proposed take-none strategy could change the level estimates by more than 1% for about a quarter of the domains. This could lead to breaks in the series and future revision of previous estimates. The absolute relative differences seem to be larger for estimates of totals than estimates of ratios.

Table 4: Distribution of the Absolute Relative Differences (%) of the Estimates Using the Proposed Take-None Strategy at the Canada Level

Variable	*P95*	*Q3*	*Median*	*Q1*	*P5*
A_AWE	4,36	0,96	0,27	0,02	0,00
H_EMPL	17,85	2,59	0,59	0,05	0,00
S_EMPL	18,27	4,08	1,12	0,09	0,00
A_AWH	2,63	0,60	0,17	0,02	0,00

Table 5: Distribution of the Absolute Relative Differences (%) of the Estimates Using the Proposed Take-None Strategy at the Provincial Level

Variable	*P95*	*Q3*	*Median*	*Q1*	*P5*
A_AWE	6,23	1,30	0,31	0,01	0,00
H_EMPL	23,28	3,06	0,67	0,04	0,00
S_EMPL	28,94	5,73	1,31	0,06	0,00
A_AWH	4,61	0,96	0,24	0,01	0,00

4.2.5 Quality Indicator

Tables 6 and 7 show the proportion of NAICS3 domains that have an excellent quality indicator (coefficient of variation (CV) less or equal to 5%) and those which are publishable (CV less or equal to 35%), both at the Canada and provincial levels, for the same variables of interest as in section 4.2.3. For this study, the CV has been calculated using the mean square error which was derived from the variance of the estimate using the current take-none strategy and the square of the difference between the estimate using the proposed take-none strategy and the one using the current take-none strategy. The reason for not using the variance of the estimate using the proposed take-none strategy is that, since the actual sample size available for the simulation was smaller than what would have been required, the calculated variance were likely to be larger than it should have been.

Using the proposed take-none strategy would result in a decrease in the proportion of domains deemed excellent. However, the proposed take-none strategy would have little to no effect on the number of publishable domains.

Table 6: Proportion of Domains Falling in Excellent or Publishable Quality Indicators at the Canada Level

		Take-none strategy	
Variable	*Quality indicator (CV)*	*Current*	*Proposed*
A_AWE	Excellent (≤5%)	85.0%	81.2%
	Publishable (≤35%)	98.9%	98.7%
H_EMPL	Excellent (≤5%)	29.9%	26.6%
	Publishable (≤35%)	92.8%	90.0%
S_EMPL	Excellent (≤5%)	5.8%	5.4%
	Publishable (≤35%)	93.4%	91.4%
A_AWH	Excellent (≤5%)	92.5%	90.7%
	Publishable (≤35%)	99.0%	98.7%

Table 7: Proportion of Domains Falling in Excellent or Publishable Quality Indicators at the Provincial Level

		Take-none strategy	
Variable	*Quality indicator (CV)*	*Current*	*Proposed*
A_AWE	Excellent (≤5%)	40.8%	38.7%
	Publishable (≤35%)	58.1%	58.1%
H_EMPL	Excellent (≤5%)	16.6%	14.2%
	Publishable (≤35%)	54.9%	53.2%
S_EMPL	Excellent (≤5%)	5.3%	5.0%
	Publishable (≤35%)	49.4%	48.2%
A_AWH	Excellent (≤5%)	46.8%	44.5%
	Publishable (≤35%)	58.1%	58.1%

5. Conclusion

The SEPH had features that managed response burden, but two new proposals that further lower response burden have recently been tested. The first proposal, which was to use the microstrata method to control sample selection, proved to be an effective way to reduce the response burden of the establishments and has been implemented in the survey production since August 2010. The second proposal, which was to increase the size of the take-none strata, also reduced the response burden, but it came with some risks. It has not been decided yet whether this proposal will be implemented in the survey production.

The next step in improving the response burden would be to find a strategy for optimizing the sampling and collection processes at the enterprise level, which is the actual survey respondent, considering that the establishment is the sampling unit.

Acknowledgements

The author would like to thank the reviewers for their useful comments.

References

Morin, M. (2010), “Rapport méthodologique Enquête sur l’emploi, la rémunération et les heures de travail (EERH)”, internal publication

Rivière, P. (2001), "Coordinating Samples Using the Microstrata Methodology", Proceedings of Statistics Canada Symposium Session 8