

CHALLENGES AND GAINS IN USING ADMINISTRATIVE DATA IN THE UNIFIED ENTERPRISE SURVEY

Chi Wai Yeung¹

ABSTRACT

The Unified Enterprise Survey (UES) is a collection of annual business surveys with unified approaches to sampling, collection, edit and imputation, as well as estimation. Starting in reference year 2003, the tax replacement initiative was implemented to lower response burden and costs of the survey programs. For some simple, incorporated businesses that are selected in sample, Statistics Canada will not contact them directly but will rather use administrative data to estimate their business activities. With the continuous improvements in the quality of the administrative data and the opportunity of a complete redesign, Statistics Canada is exploring ways to maximize the use of administrative data in the UES design. This paper examines one possible new design and compares the new design with the current design in an empirical study.

KEY WORDS: Administrative Data, Empirical Study, Survey Redesign, Tax Replacement

RÉSUMÉ

L'Enquête unifiée auprès des entreprises (EUE) est un ensemble d'enquêtes-entreprises annuelles qui préconisent la même approche en ce qui concerne l'échantillonnage, la collecte de données, la vérification, l'imputation et l'estimation. Débutant lors de l'année de référence 2003, l'initiative de remplacement par des données fiscales a été mise en oeuvre afin de réduire le fardeau de réponse et les coûts afférents aux programmes d'enquêtes. Pour certaines entreprises simples et constituées en société qui ont été sélectionnées dans l'échantillon, Statistique Canada ne les contactera pas directement et utilisera plutôt des données administratives pour estimer leur activité commerciale. En bénéficiant de l'amélioration continue de la qualité des données administratives et en ayant l'opportunité de procéder à une refonte complète de la conception de l'enquête, Statistique Canada examine des façons de maximiser l'utilisation des données administratives dans le plan d'échantillonnage de l'EUE. Cet article fait l'objet d'une comparaison entre un nouveau plan potentiel de l'EUE et le plan actuel, par l'entremise d'une étude empirique.

MOTS CLES: Données administratives; étude empirique; remaniement d'enquête; remplacement par données fiscales.

1. INTRODUCTION

The Unified Enterprise Survey is a collection of annual business surveys that serves as a vehicle for producing annual estimates for many industries at a variety of geographical levels. Among other things, its data are used in the calculations of Gross Domestic Product (GDP) and in allocation formula of the Harmonized Sale Tax (HST). In an ongoing effort to reduce response burden and costs of the survey program, the tax replacement project was initiated: instead of surveying some of the sampled businesses directly, administrative data (also known as tax data) is used to estimate their activities.

In order for a unit to be eligible for tax replacement, it must be a simple enterprise. Basically, a simple enterprise is one that has a simple structure on Statistics Canada's Business Register (BR). More specifically, it is not multi-establishment, multi-province, multi-industry, nor multi-legal. All units that are not simple are considered complex and are not eligible for tax replacement. Without directly contacting the complex businesses, the allocation of administrative data to the individual establishments, if possible at all, may not yield figures with the best data quality.

In addition, the business has to be an incorporated business which files tax using the T2 form to be eligible for tax replacement in the UES. Canada Revenue Agency (CRA) gives all incorporated businesses a Business Number (BN) thus making them easier to be linked to the BR. Also, incorporations file tax using the standardized General Index of Financial Information (GIFI) and Statistics Canada has mapped variables on the GIFI to the survey variables through the Chart of

¹ Chi Wai Yeung, Business Survey Method Division, Statistics Canada, Ottawa, Canada, K1A 0T6, chiwai.yeung@statcan.gc.ca

Account (COA). This makes it more straightforward to use administrative data of incorporations. So in summary the eligible units for UES tax replacements are simple incorporated businesses.

The UES collects both financial variables (e.g. revenues, expenses, and salary and wages, etc.) and characteristic variables (e.g. distribution by types of customers, gross leasable area, etc.) Since administrative data covers most financial variables, units selected for tax replacement will not be asked for values of these variables. Their characteristics variables are obtained either through a characteristics survey (a scaled down version of the full questionnaire) or through imputation.

In this paper, section 2 presents an overview of the current UES sampling design and in particular how tax replacement fits in the design. Section 3 presents one new proposal that seeks to maximize the use of administrative data while section 4 shows some challenges and workarounds of implementing the proposal to the UES. And finally, section 5 presents an empirical study whose objective is to compare the proposal with the current design.

2. CURRENT UES SAMPLE DESIGN

2.1 Current Sampling Strategy

The UES employs a stratified simple random sampling without replacement design. Given its mandate to “measure final sales of goods and services accurately, on an annual basis by province, and in sufficient industry details”, the stratification variable is first formed by cross-classifying province and industrial groups (using the North American Industrial Classification System or NAICS). This is known as a CELL. Then within each CELL, a take-all stratum, two take-some strata, and a take-none stratum are defined based on some SIZE variable. The take-some stratum associated with larger values of the SIZE variable possesses a greater sampling fraction and consequently, for each CELL, units from this stratum have a higher inclusion probability (π_i). As the names implied, all units within the take-all stratum have π_i equal to 1, and units in the take-none stratum have $\pi_i = 0$. Units in the take-none are not sampled and are estimated entirely based on administrative data.

The take-none units are treated differently from the tax replacement units. Since the focus of this paper is on tax replacement, the take-none will not be further discussed.

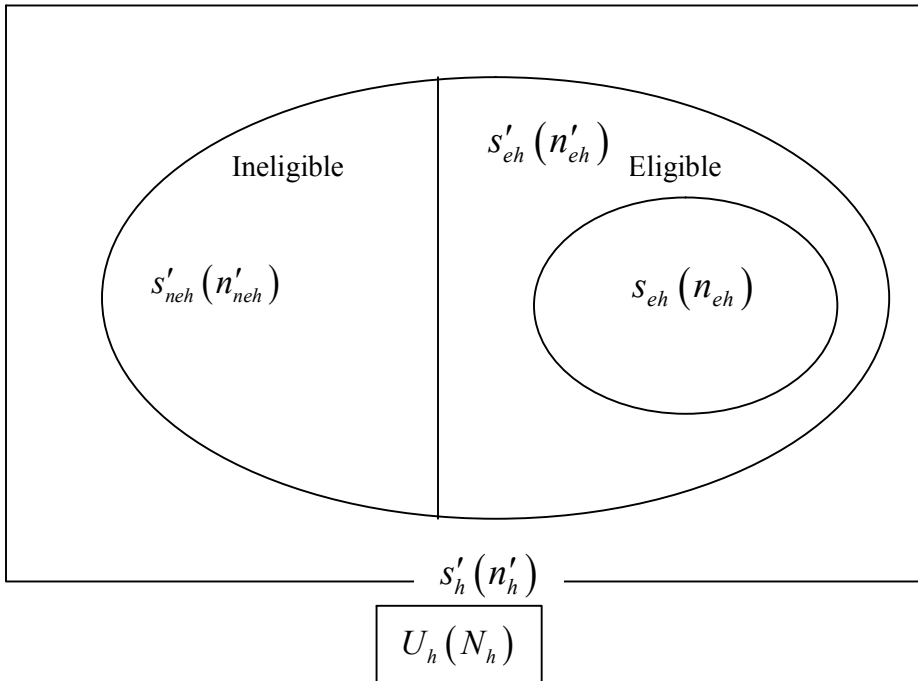
2.2 The Tax Replacement

As mentioned above in the introduction, the tax replacement units are units selected in sample but to which Statistics Canada chooses not to send questionnaires. Among the eligible businesses selected in the main sample, the UES selects a sub-sample to be sent questionnaire and tax replaces the rest. More formally:

- A. The population consists of H mutually exclusive and exhaustive strata U_h (CELL x SIZE) of size N_h , $h=1, \dots, H$. From each stratum h (excluding the take-none), a simple random sample without replacement, s'_h , of size n'_h , is selected. This is the main sample
- B. The units selected in s'_h are stratified into two strata: the stratum, s'_{eh} , of eligible units and the stratum, s'_{neh} , containing the ineligible units. $s'_h = s'_{eh} \cup s'_{neh}$. Suppose there are n'_{eh} eligible units and n'_{neh} ineligible units in s'_h . That is, $n'_h = n'_{eh} + n'_{neh}$.
- C. From s'_{eh} , a simple random sample without replacement s_{eh} , of size n_{eh} is selected. In practice the ratio n_{eh} / n'_{eh} is fixed. The units in s_{eh} go for field collection whereas the units in $s'_{eh} \setminus s_{eh}$ are tax replaced. Also, the units in s'_{neh} go for field collection.

Here is a diagram to illustrate the concepts:

Figure 1 - Current tax replacement in stratum h (CELL x SIZE)



To give an idea to the extent of tax replacement, table 1 below shows the counts of the different types of units from reference year 2008. In particular, note that the tax replacement portion consists of more than half of the eligible portion $((n'_{eh} - n_{eh}) / n'_{eh} > 50\%)$.

Table 1: Breakdown of the different types of units, reference year 2008

Type of Units	Counts
In sample, eligible units not tax replaced $(\sum_h n_{eh})$	37,446
In sample, tax replacement units $(\sum_h (n'_{eh} - n_{eh}))$	40,800
In sample but ineligible units $(\sum_h n'_{neh})$	24,936
Eligible units not in sample (but have administrative data)	256,869

2.3 Drawbacks of Tax Replacement in Its Current Form

When tax replacement was initiated back in reference year 2003, Statistics Canada did not have the resources to completely redesign the system to optimally use all administrative data. Any changes had to fit into the existing design. As a result, there are two drawbacks of tax replacement in its current form:

- A. The tax replacement portion is essentially a second phase sample of a two-phase design. The sizes n'_{eh} and n'_{neh} are random variables whose values depend on the selected first-phase sample s'_h . In other words, before selecting s'_h , one cannot predict with certainty the values of n'_{eh} and n'_{neh} and hence the value of n_{eh} . The random nature of the sample sizes may have the effect of increasing the variance of the resulting estimators.
- B. Administrative data is available for all eligible businesses, including those outside of the first phase main sample. The current use of tax data is limited to the second phase sub-sample $s'_{eh} \setminus s_{eh}$. Therefore, the current design is not making the most use of all available administrative data. In reference year 2008, this represents 256,869 units.

It is worth mentioning that the quality of the administrative data was not as good in reference year 2003 as that of today. So the current design was actually a good compromise. But given the continuous improvement in tax data quality and the opportunity of a UES redesign, several new sampling design options are being studied, one of which will be presented in the rest of this article.

3. ONE PROPOSED DESIGN

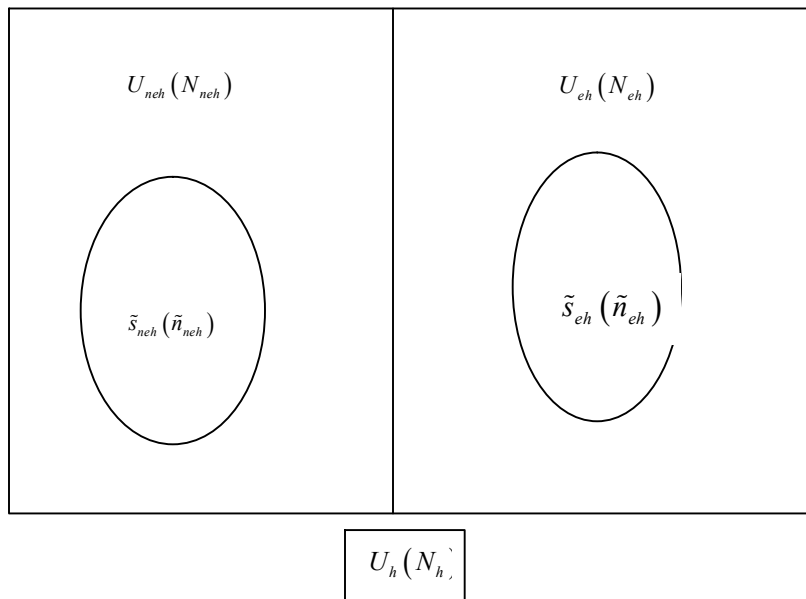
Since the UES objectives after the redesign still need to be finalized with the data users, the sampling design presented here may not be the most optimal with respect to the final goals and can very well be different from the final design. However, if publishing financial estimates remains a major mandate, the sampling design presented here is worth considering. It achieves one major goal of the redesign, namely maximizing the use of tax data.

The sampling design follows the proposal by Haziza and Kuromi (2007) and can be summarized as follow:

- A. The stratum h is first stratified into two strata: the strata, U_{eh} , of eligible units and the strata, U_{neh} , containing the ineligible units. $U_h = U_{eh} \cup U_{neh}$. Let N_{eh} be the size of U_{eh} , and N_{neh} be the size of U_{neh} . That is, $N_h = N_{eh} + N_{neh}$.
- B. From U_{eh} , a simple random sample without replacement \tilde{s}_{eh} , of size \tilde{n}_{eh} is selected. The units in \tilde{s}_{eh} go for field collection, whereas the units in $U_{eh} \setminus \tilde{s}_{eh}$ are tax replaced. Similarly, from U_{neh} , a simple random sample without replacement \tilde{s}_{neh} , of size \tilde{n}_{neh} is selected and these \tilde{n}_{neh} units go for field collection. Recall that no tax replacement takes place in U_{neh} .

Figure 2 below is a diagram to illustrate the concepts:

Figure 2: Proposed tax replacement in stratum h (CELL x SIZE)



The proposed design offers the following advantages:

- A. It is a single phase sampling design, which will make the estimation procedures somewhat simpler
- B. The sizes \tilde{n}_{eh} and \tilde{n}_{neh} are determined prior to sampling
- C. Full use of tax data is now made, since it is performed in the set $U_{eh} \setminus \tilde{s}_{eh}$. As a result, one can expect to improve the efficiency of the estimators if the tax variables are highly correlated with survey variables

3.1 Optimal Allocation under the Proposed Design

This section summarizes the optimal allocation for \tilde{n}_{eh} and \tilde{n}_{neh} from Haziza and Kuromi's paper. Given the need to be as cost efficient as possible, the problem of optimal allocation is very important.

Let y_i denotes survey data for unit i and x_i denotes tax data for unit i , then under the model of direct tax replacement:

$$y_i = x_i + \varepsilon_i \quad (1)$$

An estimator for $Y = \sum_{i \in U} y_i$ is given by:

$$\hat{Y} = \sum_{h=1}^L N_{neh} \bar{y}_{neh} + \sum_{h=1}^L N_{eh} [\bar{y}_{eh} + \bar{X}_{eh} - \bar{x}_{eh}], \quad (2)$$

where $\bar{y}_{neh} = \frac{1}{\tilde{n}_{neh}} \sum_{i \in \tilde{s}_{neh}} y_i$, $\bar{y}_{eh} = \frac{1}{\tilde{n}_{eh}} \sum_{i \in \tilde{s}_{eh}} y_i$, $\bar{x}_{eh} = \frac{1}{\tilde{n}_{eh}} \sum_{i \in \tilde{s}_{eh}} x_i$, and $\bar{X}_{eh} = \frac{1}{N_{eh}} \sum_{i \in U_{eh}} x_i$. Note the estimator in (2) consists of using a Horvitz-Thompson estimator, $N_{neh} \bar{y}_{neh}$, to estimate the individual strata totals corresponding to the ineligible units and using a difference estimator, $N_{eh} [\bar{y}_{eh} + \bar{X}_{eh} - \bar{x}_{eh}]$, for the strata corresponding to the eligible units. The difference estimator naturally comes out because of the model of direct tax replacement.

The variance of \hat{Y} in (2) is given by:

$$V_p(\hat{Y}) = \sum_{h=1}^L N_{neh}^2 \left(1 - \frac{\tilde{n}_{neh}}{N_{neh}}\right) \frac{S_{neyh}^2}{\tilde{n}_{neh}} + \sum_{h=1}^L N_{eh}^2 \left(1 - \frac{\tilde{n}_{eh}}{N_{eh}}\right) \frac{S_{edh}^2}{\tilde{n}_{eh}}, \quad (3)$$

where $S_{neyh}^2 = \frac{1}{N_{neh} - 1} \sum_{i \in U_{neh}} (y_i - \bar{Y}_{neh})^2$ with $\bar{Y}_{neh} = \frac{1}{N_{neh}} \sum_{i \in U_{neh}} y_i$ and $S_{edh}^2 = \frac{1}{N_{eh} - 1} \sum_{i \in U_{eh}} (d_i - \bar{D}_{eh})^2$ with $d_i = y_i - x_i$ and $\bar{D}_{eh} = \frac{1}{N_{eh}} \sum_{i \in U_{eh}} d_i$.

If the cost function is of linear form:

$$C = c_0 + \sum_{h=1}^L c_h \tilde{n}_{neh} + \sum_{h=1}^L c_h \tilde{n}_{eh}, \quad (4)$$

where c_0 is a fixed overhead cost and c_h denotes the cost of surveying a unit in the field. At this point, one is seeking values \tilde{n}_{neh} and \tilde{n}_{eh} that minimize the variance (3) given the cost (4). After some algebra, one obtains

$$\tilde{n}_{neh} = n \frac{N_{neh} S_{neyh} / \sqrt{c_h}}{\sum_{h=1}^L N_{neh} S_{neyh} / \sqrt{c_h} + \sum_{h=1}^L N_{eh} S_{edh} / \sqrt{c_h}}$$

and

$$\tilde{n}_{eh} = n \frac{N_{eh} S_{edh} / \sqrt{c_h}}{\sum_{h=1}^L N_{neh} S_{neyh} / \sqrt{c_h} + \sum_{h=1}^L N_{eh} S_{edh} / \sqrt{c_h}}.$$

4. CHALLENGES AND WORKAROUNDS

When one tries to implement the proposal to the UES, one quickly realizes that two adjustments are needed.

4.1 Estimating S_{edh}

S_{edh} are required to derive \tilde{n}_{eh} and \tilde{n}_{neh} . However, some surveys within the UES are currently using 100% tax replacement. In other words, the ratio $n_{eh} / n'_{eh} = 0\%$, meaning y_i is missing from all eligible units by design and there is no way to calculate the d_i and hence S_{edh} .

Even for surveys with $n_{eh} / n'_{eh} > 0\%$, there are also units in s_{eh} that are non-respondents in either y_i or x_i or both. As a result, only a small portion of units has both y_i and x_i . Imputation is needed to estimate S_{edh} . The imputation is carried out in stages. First of all, if there are two or more units with both reported y_i and x_i within CELL x SIZE, the estimate of S_{edh} (\hat{S}_{edh}) will be calculated based on these units. Otherwise, the SIZE dimension is collapsed. If two or more units exist at the CELL level, S_{edh} will be estimated by incorporating the weight:

$$\hat{S}_{edh,CELL} = \sqrt{\frac{1}{\sum_{i \in CELL} w_i - 1} \sum_{i \in CELL} w_i (y_i - x_i)^2} \text{ where } w_i = 1/\pi_i \text{ is the design weight}$$

This $\hat{S}_{edh,CELL}$ will be used for those CELL x SIZE without \hat{S}_{edh} . This process is repeated using NAICS as the third level and finally the entire survey as the fourth level.

The author acknowledges that the threshold of having only two units may lead to unstable \hat{S}_{edh} . But given the little amount of data at the onset, it is used to get \hat{S}_{edh} at the lowest level possible. Table 2 shows the level in which \hat{S}_{edh} are derived for the two surveys used in the empirical study below.

Table 2: The level in which \hat{S}_{edh} is derived

Level	Wholesale	Retail
CELL x SIZE	83	120
CELL	111	228
NAICS	16	28
Entire Survey	2	0

4.2. Modification - the Individual Cell Approach

The allocation formula for \tilde{n}_{eh} and \tilde{n}_{neh} minimize the proposed variance (3) at the survey level. However, given the UES mandate to publish by CELL, the allocation may not be optimal at the CELL level. Also, there is no evidence that c_h are different across stratum, so the formula can be simplified slightly to become:

$$\tilde{n}_{neh} = n_h \frac{N_{neh} S_{neyh}}{N_{neh} S_{neyh} + N_{eh} S_{edh}} \quad (5)$$

$$\tilde{n}_{eh} = n_h \frac{N_{eh} S_{edh}}{N_{neh} S_{neyh} + N_{eh} S_{edh}} \quad (6)$$

where n_h are the sample size of CELL x SIZE generated by the existing Lavallée-Hidiroglou algorithm (1988), using a power allocation methodology.

5. EMPIRICAL STUDY

An empirical study was carried out using 2006 data from the Wholesale sector and the Retail sector. These two sectors were chosen since the ratio n_{eh} / n'_{eh} is around 45%², making it possible to calculate \hat{S}_{edh} . The goal is to evaluate the gains in efficiencies of the new design. This information will be useful to compare the different redesign proposals.

The empirical study consists of calculating \tilde{n}_{eh} and \tilde{n}_{neh} using (5) and (6) above as well as calculating the variance of the estimator, both at the CELL level and the Survey x SIZE level. We are interested in the Survey x SIZE level since the y_i and hence the variability should be larger in the large take-some, so one would expect the gain in efficiency to be more significant. The y_i used is *Total Operating Revenue* while the x_i is the corresponding COA variable. Since n_h and hence $n = \sum_{h=1}^H n_h$ are obtained directly from the existing sample design, lower variances in the proposed design can be considered empirical evidence that the proposal is more efficient.

Table 3 below shows the breakdown between \tilde{n}_{eh} and \tilde{n}_{neh} by the SIZE group:

Table 3: Breakdown of \tilde{n}_{eh} and \tilde{n}_{neh} by survey x SIZE

	N_{neh}	\tilde{n}_{neh}	N_{eh}	\tilde{n}_{eh}
Retail small take-some	14,112	629	42,306	528
Retail large take-some	3,277	637	11,154	569
Wholesale small take-some	2,684	171	12,519	317
Wholesale large take-some	723	202	2,764	291

Some insights can be gained by converting the table above to percentages ($N_h = N_{neh} + N_{eh}$; $n_h = \tilde{n}_{eh} + \tilde{n}_{neh}$):

Table 4: Percentage of \tilde{n}_{neh}/n_h and \tilde{n}_{eh}/n_h

	N_{neh}/N_h	\tilde{n}_{neh}/n_h	N_{eh}/N_h	\tilde{n}_{eh}/n_h
Retail small take-some	0.250	0.544	0.750	0.456
Retail large take-some	0.227	0.528	0.773	0.472
Wholesale small take-some	0.177	0.350	0.823	0.650
Wholesale large take-some	0.207	0.410	0.793	0.590

One thing to note from table 4 is that there are more units allocated to the ineligible portion compared to proportional allocation, implying that in the allocation formula (5) and (6), S_{edh} are indeed smaller than S_{neyh} . This may be a sign of gain in efficiency.

Table 5 below shows the survey variances by SIZE group

Table 5: Comparing the total variance by SIZE

	Current (10^{18})	Proposal (10^{18})
Retail small take-some	3.73	4.84
Retail large take-some	4.26	2.14
Total	7.99	6.98
Wholesale small take-some	12.91	6.88
Wholesale large take-some	12.16	1.39
Total	25.07	8.27

² For the Wholesale and Retail surveys, some CELLS are not tax replaced (e.g. Wholesales Agents and Brokers). This is something to keep in mind in the empirical study.

Even though the proposed design shows lower variance in both surveys, the improvement in Wholesale is more substantial. That is because Wholesale has much larger variance under the current estimator. Typically firms in the Wholesale sector have larger y_i , so variance under the Horvitz-Thompson estimator is larger. On the other hand the proposed design and its variance estimator use d_i , which are less sensitive to the difference in industry.

It may be strange that the variances are higher in the small take-somes than the large take-somes under the proposal. The author notes that this is an indirect result of the method used to derive \hat{S}_{edh} . Table 2 shows that the majority of \hat{S}_{edh} are derived by collapsing SIZE so the term should be the same across the two take-some strata. According to the variance formula (3), the remaining terms are \tilde{n}_{neh} , \tilde{n}_{eh} , S_{neyh}^2 , N_{neh} , and N_{eh} . Table 3 shows that N_{neh} , and N_{eh} are much larger in the small take-somes. So even though $S_{neyh}^2 / \tilde{n}_{neh}$ are smaller in the small take-somes, they are not enough to compensate for the difference in N_{neh} , and N_{eh} .

Table 6 below shows the results at the CELL level. In particular, it shows the number of CELLS in which the proposed design and estimator has a lower variance than the current design and estimator.

Table 6: Number of CELLS in which the proposal has lower variance

	Wholesale	Retail
No. of CELL with lower variance	91	165
Total No. of CELL	115	199
Percentage	79%	83%

At the CELL level, there is indeed empirical evidence showing gains in efficiency. To study the distribution of the differences, histograms of $\log(\hat{V}_{proposal}) - \log(\hat{V}_{current})$ are plotted, where $\hat{V}_{proposal}$ is similar to (3) except it is at the CELL level while $\hat{V}_{current}$ is the standard variance estimator of a Horvitz-Thompson estimator. If $\hat{V}_{proposal} = 0$ then $\log(\hat{V}_{proposal})$ is set to 0 for ease of comparison.

Figure 3 - Histogram of $\log(\hat{V}_{proposal}) - \log(\hat{V}_{current})$ for the CELL in Wholesale

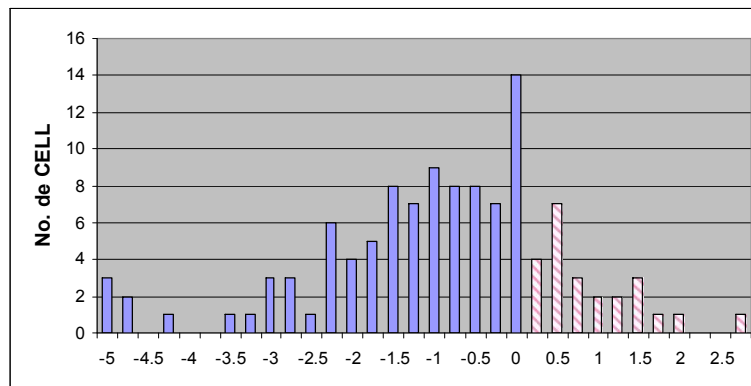
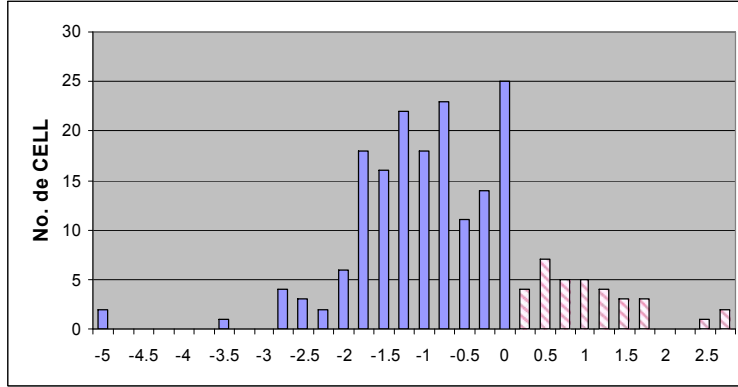


Figure 4 - Histogram of $\log(\hat{V}_{proposal}) - \log(\hat{V}_{current})$ for the CELL in Retail



The author would make two observations:

- A. There are several CELLS with $\log(\hat{V}_{proposal}) - \log(\hat{V}_{current})$ exactly equal to 0. They are CELLS that are not eligible for tax replacement (e.g. Wholesales Agents and Brokers) so the variance should be the same between the proposal and the current design.
- B. There are a few CELLS with extremely negative values (e.g. less than -4). The majority of them are CELLS in which $\hat{V}_{proposal}=0$ so $\log(\hat{V}_{proposal})$ are set to 0. The reason $\hat{V}_{proposal}$ can be 0 is because \hat{S}_{edh}^2 is estimated to be 0 and \tilde{n}_{neh} from the allocation formula is equal or greater than N_{neh} . So the finite population correction will make the variance in the ineligible part zero while \hat{S}_{edh}^2 will make the eligible part zero.

5. CONCLUSION

The empirical exercise shows that some workarounds are needed to implement the proposed design in the UES context. In particular, producing estimates of S_{edh} proved to be challenging given the current tax replacement design and non respondents. Since some \hat{S}_{edh}^2 are zero in the set up of using only two respondents as a threshold, more robust estimates of S_{edh} are desirable. The author can also foresee another issue with the proposed design. Theoretically, if survey data and tax data are highly correlated, \tilde{n}_{eh} will be small compared to \tilde{n}_{neh} , which will add to the difficulty of estimating S_{edh} in subsequent years.

The empirical study shows that there is some evidence of gains in efficiency. But since Statistics Canada has yet to finalize the survey objectives with the data users, this proposal may not be the most suitable given the new objectives. For example, if the focus after the redesign is on the characteristics variables, the proposal may be less than ideal. It is in the best interest of Statistics Canada to clearly establish its objectives before finalizing the new design.

ACKNOWLEDGEMENTS

The author would like to thank Claude Turmelle, Gordon Kuromi, and Patrice Mathieu for their useful comments in improving this paper.

REFERENCES

Haziza, D. and Kuromi, G. (2007). "Sampling Design for the Unified Enterprise Survey". *Statistics Canada Internal Document*

Lavallée, P. and Hidioglou M.A. (1988). "On the Stratification of Skewed Populations". *Survey Methodology*, **14**, 33-43.