DEALING WITH THE PROBLEM OF COMBINED REPORTS AT THE SAMPLING DESIGN STAGE FOR THE WORKPLACE AND EMPLOYEE SURVEY

Cynthia Bocci[1] and Jean-François Beaumont[2]

ABSTRACT

The Workplace and Employee Survey is a longitudinal survey that collects information on employers (at the location level) and on employees working in the selected locations. Enterprises having more than one location, referred to as multi-location employers, are often unfortunately not able to report for a chosen location and report combined information which is simply unusable or imputed in many cases. These combined reports have the effect of introducing measurement errors and thus reducing data quality. In this article, we discuss various options to select fewer multi-locations in the sample in order to increase data quality. We also investigate how much increase of variance is acceptable in order to reduce the number of combined reports.

KEY WORDS: Coefficient of variation, Conditional bias, Multi-location, Stratification.

RÉSUMÉ

L'enquête sur le milieu de travail et les employés est une enquête longitudinale qui recueille des informations auprès des employeurs (au niveau de l'emplacement) et des employés travaillant dans les emplacements sélectionnés. Les entreprises qui ont plus d'un emplacement, ce qu'on appelle des employeurs multi-emplacements, ont malheureusement souvent de la difficulté à rapporter les données pour un emplacement choisi. Ils fournissent donc un rapport combiné, lequel est tout simplement inutilisable ou imputé dans plusieurs cas. Ces rapports combinés ont pour effet d'introduire des erreurs de mesure et ainsi de réduire la qualité des données. Dans cet article, on discute de plusieurs options pour sélectionner moins de multi-emplacements dans l'échantillon afin d'accroître la qualité des données. On étudie également à quel point une augmentation de la variance est acceptable afin de réduire le nombre de rapports combinés.

MOTS CLÉS : Biais conditionnel; coefficients de variation; multi-emplacements; stratification.

1. INTRODUCTION

The Workplace and Employee Survey (WES) is an annual survey with two components. The first component consists of a stratified simple random sample of employers at the location level. Data is collected from a contact person for each chosen location to measure competitiveness, innovation, technology use and management of human resources. The second component consists of a systematic random sample of employees from each of the selected employers. Data from employees is collected to measure technology use, training and stability of employment and income.

The problem of interest pertains only to the employer portion of the survey. Consider an employer (location) which belongs to an enterprise with many inscope locations. Define such a location as a multi-location. Only data from the selected location is of interest in the survey. Unfortunately, in the case of a multi-location, the contact person frequently is unable to provide data on only the chosen location and submits a combined report whereby the reported values represent many locations. From such a combined report it is often difficult to extract information for the selected location.

[1] Cynthia Bocci, Statistics Canada, Business Survey Methods Division, R.H.Coats Bldg. 11th Floor, 150 Tunney's Pasture Driveway, Ottawa, Ontario K1A 0T6, Cynthia.Bocci@statcan.ca
[2] Jean-François Beaumont, Statistics Canada, Business Survey Methods Division, R.H.Coats Bldg. 11th Floor, 150 Tunney's Pasture Driveway, Ottawa, Ontario K1A 0T6, Jean-Francois.Beaumont@statcan.ca

Doing so introduces errors and can reduce the quality of the data. Many times, these combined reports are simply thrown out since it is impossible to derive any useful information from them.

One goal for the WES redesign is to reduce the number of these potential combined reports. This can be achieved through stratification by finding a scheme that reduces the number of multi-locations selected while producing estimates of similar quality. In addition, the impact of the design weights resulting from such a stratification scheme is investigated in an effort to address the problem of stratum jumpers (i.e. miss-classified units on the frame) later on in the estimation phase. Throughout this document, multi-locations are referred to as multis whereas locations that are not part of an enterprise with several inscope workplaces are referred to as singles.

The study aims to find a compromise between the reduction in the number of multi-locations and the increase in variance that could result. In this article, we will compare several measures for two different allocation schemes obtained by using the 1999 WES population and sample data files.

2. STRATA AND ALLOCATION SCHEMES

We decided to maintain a one-stage stratified simple random sample of locations (employers) and construct various strata using combinations of variables available on the frame. Initially, we considered several different strata definitions and then narrowed our studies to the most promising one. In the article, we restrict our discussion to the current stratification and one alternate stratification scheme.

2.1 Stratification variables

The stratification variables under consideration are industry, region, size of workplace based on the number of employees and an indicator of whether or not a workplace is part of a multi-location. Specifically, there are 14 industries and 6 regions grouping the provinces. The current stratification is defined by *industry/region/size_mb* where size_mb is a variable grouping size of a workplace into 3 categories. The boundaries delimiting the 3 categories are industry/region specific and are determined by using a model-based approach to stratification (Patak et al., 2000). There are 252 strata in the current scheme.

The proposed stratification scheme is defined by *industry/region/size/multiflag* where size is a categorical variable grouping the number of employees into 4 categories and multiflag is an indicator variable identifying multi-locations. The 4 size categories are fixed and correspond to domains of interest for the analysts. This stratification scheme resulted in 648 non-empty strata in the 1999 WES population.

2.2 Allocation scheme

Under the current scheme, we used Neyman allocation within an industry/region with equal coefficients of variation (CV) at this level to yield a sample size of approximately 7400. The minimum number of units sampled per stratum was set to 10 although no collapsing in this study was done if there were fewer than 10 population units in the stratum.

Under the proposed scheme, we decided that large employers within a specific industry should be sampled with certainty due to the significant impact of large employers in the workplace. Employers with the number of employees greater than an industry specific lower bound were considered must-take and became a take-all (TA) group. We used Neyman allocation as described above for the singles (a minimum of 10 units are sampled where possible) and chose 3 multis at random within each multi-type stratum.

2.3 Minimum sampling fractions

In an effort to better control the sampling weights and to reduce the effects of stratum jumpers in the estimation phase, we first alter the minimum sampling fraction, $f_{\min}$, for the strata of single-type locations in the proposed stratification scheme. The minimum number of sampled units in a single-type stratum h is defined as $n_{\min_h} = \text{maximum}(10, f_{\min} * N_h)$ where N_h is the population size of stratum h. In our study, we allow $f_{\min}$ to vary from .001 to .006. The value of .003 produces satisfying results according to certain measures described in Section 4. Next, we fix the single-type minimum sampling

fraction at .003 and allow the minimum sampling fraction for the multi-type strata to vary slightly. The minimum number of sampled units in a multi-type stratum is equal to $n_{\min_h} = \max imum(3, f_{\min} * N_h)$.

3. POPULATION VARIABLES

It is necessary to evaluate the effects of a given stratification scheme on key variables of interest. This section describes the variables which will be considered when evaluating the effectiveness of a given scheme.

The number of employees for a given employer is available for all units in the WES population through administrative information. Nonetheless, sampled employers are asked to provide the number of employees on the survey. This is a key variable and is estimated at the industry/region level. For the purposes of this study, effects on estimation of certain financial survey variables are also of interest. In particular, we investigate the effects of stratification schemes on five financial variables by imputing these variables for the unsampled population and keeping their values for the sample. Using 1999 WES sample data, we impute using nearest-neighbour imputation method with respect to the number of employees. Once a donor from the sample is identified, all five financial variables are simultaneously imputed using the chosen donor. The imputed variables are: benefits, training, revenue, expenditures and gross payroll. In addition, a random exponential variable, denoted *ysim*, is created for all population units. This variable acts as a control to observe the impact of the stratification scheme on a variable uncorrelated with the number of employees.

4. MEASURES OF COMPARISON

Several different measures of comparison are used to compare the current and proposed stratification schemes. In accordance with the major goal of this study, our first concern was to verify the number of multi-locations sampled as a result of each stratification scheme. Next, the expected CVs at the industry/region level are calculated for all the population variables of interest described in the previous section.

Finally, we construct a measure of the impact of an employer within its stratum. The bias of the Horvitz-Thompson estimator of a total for the variable *y*, conditional on the samples *s* containing unit *i* in stratum *h* is given by

$$B_{hi}(\hat{t}_y^{HT} \mid (hi) \in s) = (w_h - 1)\frac{N_h}{N_h - 1}(y_{hi} - \overline{Y}_h)$$

where N_h is the population size of stratum h, w_h is the design weight for stratum h and $\overline{Y}_h$ is the average population value in stratum h. This conditional bias can be viewed as a measure of influence. The unit will have more influence if the design weight is large or if the value of the variable y for that unit differs a great deal from the average value in that stratum. A conditional relative bias measure, $100 * B_{hi} / \overline{Y}_{domain}$, can be defined for each unit i in stratum h belonging to a domain. This relative bias was calculated for all units in the population using industry/region for a domain.

5. RESULTS

Table 1 shows that the proposed method reduces the number of multi-locations by approximately half while improving the efficiency for estimating the number of employees. The reduction is simply due to construction of the proposed stratification scheme whereas the gain in efficiency is mainly due to the categorization of the size variable.

Table 1: Number of multis and expected CV under two schemes

Scheme	Number of strata	Percentage of TA strata	Sample size	Number of multis	Percentage of multis	Average expected CV (%) for employment at industry/region level
Proposed	648	14.2	7221	1220	16.9%	7
Current	252	7.9	7415	2402*	32.4%	9

* averaged over 100 stratified SRS random samples

Table 2 gives the distribution statistics on the average CV (%) over the 84 industry/region levels for all the variables of interest in both stratification schemes. While there is a gain in efficiency in estimating the number of employees with the proposed method, there is a small loss for the other variables with an increase in standard deviation and a large increase in the maximum CV. As expected, the random exponential variable is unaffected by the proposed scheme.

Table 2: Average CV (%) over industry/region levels with minimum, maximum and standard deviation

	CV distribution							
	Current scheme				Proposed scheme			
Variable	average	Standard deviation	minimum	maximum	average	Standard deviation	minimum	maximum
Number of employees	8.9	0.1	8.8	9.4	6.8	0.3	4.8	7.1
Total expenditures on non-wage benefits	23.2	10.9	7.0	61.9	26.3	12.5	10.8	95.4
Training expenditures	26.6	12.3	4.9	67.7	35.5	22.6	9.8	129.6
Revenue	21.3	10.6	8.0	59.0	26.5	18.1	9.8	117.7
Gross expenditures	19.4	10.8	4.7	63.2	23.2	14.5	7.2	92.4
Total gross payroll	16.8	9.0	7.3	62.0	17.1	9.2	8.1	65.6
Random exponential variable	15.8	2.3	9.7	20.4	15.4	3.6	10.0	25.3

When varying the single-type minimum sampling fraction, the results (not shown here) indicate that there is no appreciable gain other than weight control. The same is true when fixing the single-type minimum sampling fraction at .003 and varying the multi-type minimum sampling fraction.

We then calculated the maximum, 95th percentile and the standard deviation of the conditional relative bias over all units in a domain for each of the seven variables under the current and proposed stratification scheme. The improvement under the proposed scheme can be seen in the 95th percentile which is smaller under the proposed scheme in most of the 84 domains for almost all of the seven variables of interest. Table 3 shows an example in the retail and trade industry for the province of Ontario containing 99,686 population units. It demonstrates that for this particular industry/region, the 95th percentile is generally smaller for the proposed scheme than the current one. Not surprisingly, the maximum conditional relative bias is much smaller for the number of employees under the proposed stratification scheme. The substantial increase of the maximum relative conditional bias for the other variables in the proposed method is due to the multi-type strata. The selection of only 3 multis in a multi-type stratum can lead to a large sampling weight resulting in a large conditional bias.

Table 3: Relative conditional bias (in thousands) for the retail and trade industry in Ontario

	Relative conditional bias distribution					
	Current scheme			Proposed scheme		
Variable	maximum	95th percentile	standard deviation	maximum	95th percentile	standard deviation
Number of employees	2 222	152	77	640	144	74
Total expenditures on non-wage benefits	4 074	273	339	47 614	157	575
Training expenditures	8 246	364	387	38 372	296	545
Revenue	3 756	96	603	11 457	91	565
Gross expenditures	5 388	48	859	12 808	40	797
Total gross payroll	2 881	103	464	13 626	122	447
Random exponential variable	2 811	455	231	3 538	323	184

6. CONCLUSION

We used several comparison measures to study an alternative stratification scheme to the one currently used in WES. The new scheme successfully reduces the number of potential combined reports by reducing the number of multi-locations sampled. The proposed scheme is more efficient in terms of estimating the number of employees. The cost of reducing the number of multis, however, is reflected in the CVs of the other variables presented here. The proposed scheme does allow for some control of the sampling weight of singles to address in part the problem of stratum jumpers in the estimation phase.

REFERENCES

Patak, Z., Lavallée, P., Hidiroglou, M.A. (2000). The methodology of the Workplace and Employee Survey, *Proceedings of the 2nd International Conference on Establishment Surveys*, Buffalo, NY. Published by the American Statistical Association.