Statistics and Applications {ISSN 2452-7395 (online)}
Volume 16, No. 1, 2018 (New Series), pp 45-56

Sample Design for the Survey of Enterprises in Philippines

G. Hussain Choudhry
Ottawa, Ontario, Canada

Received: November 06, 2017; Revised: December 10, 2017; Accepted: December 22, 2017

Abstract

For the Survey of Enterprises in Philippines (SEP), we developed a stratified single-stage sampling design with systematic sampling of enterprises from within strata. The enterprise total employment was used as measure of size for size stratification because this was the only variable available on the Business Register (BR). The Philippines National Statistical Office (NSO) would make arrangements with the Bureau of Internal Revenue (BIR) to get tax data on key financial variables (e.g. revenue, VAT, etc.) so that these variables could be used to develop a more efficient sample design. The sample design that we developed was such that the sample of enterprises could be used to produce both the enterprise as well as the establishment estimates by treating the enterprises as clusters of establishments. There are 18 industry groups that are in-scope for the SEP. Because of very small population counts, only the national estimates would be produced for 8 out of the 18 industries. But, it would be possible to produce both the regional and the national level estimates for the remaining 10 industries. Methods used for estimation and variance estimation are also presented.

Key words: Enterprise Surveys, Sample Allocation, Sample Weighting and Estimation, Variance Estimation, Taylor Method

1. Introduction

The Philippines System of National Accounts (PSNA) consists of an integrated set of macroeconomic accounts, balance sheets, and tables based on internationally accepted concepts, classifications and accounting rules, thus, providing a comprehensive accounting framework within which economic data are compiled and presented as vital inputs to economic analysis, decision-making and policy formulation. The main objective of the investigation was to improve the quality and usefulness of the PSNA by improving the coverage and quality of the industry surveys. Among other collection issues the major issue was the high nonresponse because of refusals. The businesses were refusing to cooperate due to: the increasing number of referrals to a unit (usually head office) capable of providing the requested information; the increasing number of consolidated reporting; the growing reluctance among respondents to provide data through the many separate survey requests sent them monthly, quarterly and/or annually. Instead of addressing the nonresponse issue through repeated call-backs, we decided to implement an integrated enterprise survey approach to address the collection problems of the NSO. Using the integrated enterprise survey approach data can be collected from a single survey to produce both the enterprise level and the establishment level estimates.

Corresponding Author: G. Hussain Choudhry
E-mail: ghchoudhry@gmail.com

We considered a number of sample design options for the Survey of Enterprises in Philippines (SEP). The sample design that was implemented for the SEP is described in section 2. The field data collection operations are discussed in section 3. Finally, the methodology used for sample weighting and estimation including variance estimation with Taylor method is discussed in section 4.

2. Sample Design

2.1 Survey Objectives

The objective of the Survey of Enterprises in Philippines (SEP) was to produce reliable regional and national level estimates for each of the 18 industry groups (**2009 PSIC Section)**. In addition, the regional and national level estimates for the aggregate of all industries was required. These estimates would include both enterprise and establishment level attributes. There are 17 regions and 18 industry-groups, which would result in 306 cells in the cross-classification of region by industry. Based on the preliminary investigation (using the 2011 BR) 8 out of the 18 industries had less than 8.0 percent of establishments in the country. In fact, some of these 8 industries did not even have establishments in all 17 regions. Therefore, it was decided to produce both regional and national level estimates for the 10 larger industries, and only the national level estimates would be produced for the remaining 8 industries.

2.2 Sample Design Options

We considered three sample design options for the SEP. The sample design that was implemented was a stratified single-stage design with systematic sampling of enterprises. Sampled enterprises and all establishments belonging to the sampled enterprises were to be enumerated. Alternatively, either a sample of establishments from the selected enterprises could be enumerated or consolidated reports at the appropriate level could be obtained. The other two design options were not feasible because establishment was the sampling unit for both these options, and that would result in multiple roots of selection for the enterprises. It would not have been possible to compute the exact selection probabilities for the enterprises because the list if establishments belonging to the enterprise was not always up to date. Moreover, one of these options would also result in two sets of survey weights: one for the establishments and one for the enterprises. Although the option that we implanted was somewhat less efficient for the establishment estimates due to clustering, it has the advantage of being a simple design. It also has the advantage that it would require that the establishments be linked with their corresponding parent enterprises only for the sampled enterprises, and not for the entire frame. Since all establishments belonging to the selected enterprises will be enumerated it would reduce the establishment under-coverage. There will be no issues related to referrals because all data collection will be coordinated through the parent enterprise (head office).

2.3 Statistical Frame for the Survey of Enterprises

The sampling frame was the list of all in-scope enterprises. In order to construct the sampling frame we first defined the “target population” for the SEP. The sampling frame was then developed to provide coverage that was as close as possible to the target population. The “target population” was defined as all Enterprises in Philippines engaged in the economic activities in the 18 industry groups that were in-scope for the SEP. Moreover, the SEP would cover only those enterprises that were engaged in the formal sector. The informal sector to be excluded was defined as single unit (establishment) enterprises owned by single proprietorships, and with total employment count of less than 10.

2.4 Stratification and Sample Allocation

The region and industry was used for primary stratification, and the total employment count was used as measure of size (MOS) for size stratification within primary strata. Since enterprise was the statistical unit, and an enterprise can have establishments in multiple (two or more) regions and/or operate in multiple (two or more) industries out of the 10 larger industries such enterprises were known as **complex enterprises**. If an enterprise was not complex then it was called a ***simple*** enterprise. Before defining the primary stratification we defined a ***Must-Take*** stratum for the complex enterprises.

Must-Take Stratum

Ideally, all complex enterprises should be assigned to the ***Must-Take*** stratum but there were too many such enterprises. Since units assigned to the Must-Take strata were to be sampled with certainty, the sample size would have become too large because of large number of complex enterprises. Thus, it was not feasible to select all complex enterprises with certainty. Therefore, we used a cut-off (threshold) value of total employment to define ***Must-Take*** stratum. We used a cut-off value of 500 for total employment and found that only 278 complex enterprises got assigned to the Must-Take strata instead of close to 4,800 enterprises. Alternatively, different cut-off values could also be used for different industries. It should be noted that Complex enterprises were defined only for the 10 larger industries.

Next, we describe the primary stratification on the basis of region and industry. In order to define the primary stratification, we assigned dominant region and industry to the complex enterprises that did not get assigned to the Must-Take stratum. The dominant region and industry were assigned on the basis of total employment, i.e. the region and industry of the establishment with the largest total employment. It should be noted that the dominant region and industry were assigned simultaneously, and not sequentially.

Primary Stratification

The primary stratum was the industry itself for the 8 industries for which only the national level estimates were to be produced. For the remaining 10 industries the primary strata were defined by cross-classification of region and industry, where region and industry were respectively the dominant region and dominant industry for the complex enterprises. Thus, there were altogether 178 primary strata.

Secondary Stratification

Secondary strata were defined within each of the primary strata based on the size, which was the total employment at the enterprise level. Before defining the size strata we defined ***Take-None*** strata within each of the primary strata.

Take-None Strata

The Take-None strata within the primary strata consisted of enterprises with smaller total employment count. We defined the Take-None portion within each primary stratum using the criterion that the Take-None account for no more than 5.0 percent of total employment within the primary stratum. As the name implies no sample had to be selected from the Take-None strata. The rational for Take-None portion was to reduce the response burden and the data collection cost. Based on the partial 2011 BR we found that 40.0 percent of the enterprises were assigned to the Take-None portion. It should be pointed out that the MOS (enterprise level total employment) for the Must-Take was not taken into consideration when determining the Take-None boundary, which provided a more conservative bound for the Take-None portion.

Size Stratification and Sample Allocation

Next we determined size strata boundaries within each primary stratum, and the corresponding sample allocation. We used the enterprise total employment as the measure of size (MOS) for this purpose. We defined three size strata (one Take-All, and 2 Take-Some Strata) within each of the primary strata. We considered the following two approaches for size stratification and sample allocation.

1. **Lavallee-Hidiroglou Method:** The Lavallée-Hidiroglou (1988) method defines size stratification as well as determines the corresponding sample allocation across the size strata by minimizing the total sample size to achieve the specified CV of an auxiliary variable at the cell level. It should be noted that one of the three size strata was Take-All stratum. The L-H method satisfies the specified CV constraint at the cell level, but the CVs at the region level, the industry group level and the national aggregate level cannot be specified. Moreover, the CVs at the cell level can be controlled for one auxiliary variable only.

2. **Non-Linear Programming Approach:** The Non-Linear Programming (NLP) approach is an alternative that can be used to obtain optimum sample allocation to the size strata defined by L-H method (or any other method) to achieve the specified CVs at the cell level as well as at the margins (see Choudhry, et. al., 2012). Moreover, these CVs can be specified for multiple auxiliary variables. The SAS procedure NLP (Non-Linear Programming) can be used to minimize the total sample size (or total cost) while satisfying the specified CV constraints.

As mentioned above, the Lavallée-Hidiroglou (1988) algorithm that defines size stratification also determines the sample allocation to achieve the required CVs at the primary stratum level. The main advantage of the L-H method is that it determines the size strata boundaries as well as the sample allocation across size strata. The L-H method would control the CVs at the primary stratum level, but not at the marginal levels and the national aggregate level. Generally, these CVs would be small because the estimates are based on the aggregation of the primary strata for which the CVs were controlled.

The NLP method would require that the size strata boundaries be defined using another method, e.g. cum$\sqrt{f}$ rule (Dalenius and Hodges, 1959) or some other method. Horgan (2006) has discussed a number of methods to define size strata boundaries including the Dalenius and Hodges (1959) method and the Lavallée-Hidiroglou (1988) method. The NLP method is optimum for the given strata boundaries. It is possible that the NLP method will be less efficient than the L-H method if the strata boundaries used for NLP method are such that the resulting strata are not very homogeneous. The only advantage of the NLP method would be that the CVs can be controlled at the primary stratum level as well as at the aggregate levels for multiple variables, whereas the L-H method can only control the CVs at the primary stratum level for a single auxiliary variable. But, the main advantage of the L-H method would be that determining the size strata boundaries and allocation of sample to the size strata would be a one step process. Based on the above consideration, the L-H (1988) method for size stratification and sample allocation was implemented.

It should be noted that the variance contribution from the Must-Take stratum will be zero. But, we did include the Must-Take contribution when assigning the CVs for computing sample size and sample allocation. Therefore, in order to achieve the desired CV at the primary stratum level the CV to be specified for the L-H algorithm will be given as:

$$CV_{Specified} = \left(1 + \frac{Y_{Must-Take}}{Y_{\sim Must-Take}}\right) CV_0 \,, \tag{2.1}$$

where CV_0 is the target (or desired) CV, and $Y_{Must-Take}$ and $Y_{\sim Must-Take}$ are respectively the contributions to the primary stratum total from the Must-Take stratum and the compliment of the Must-Take stratum (i.e., enterprises belonging to the primary stratum that were not assigned to the Must-Take stratum), and $CV_{Specified}$ is the CV that would have to be specified for the L-H algorithm in order to achieve the target CV of CV_0.

3. Data Collection

3.1 Field Data Collection Operation

Generally, nonresponse would increase with response burden. Therefore, it is very important to implement field data collection procedures to minimize the response burden. The nonresponse not only increases the variances of the survey estimates but the potential nonresponse bias increases as the nonresponse level increases. This is particularly true for enterprises in the Must-Take and Take-All strata because these are large and/or complex enterprises and were selected with certainty. Therefore, the enterprises in Must-Take and Take-All strata were targeted for 100 percent response. If there was still some nonresponse in these certainty strata data were imputed using administrative sources instead of weight adjustment.

A framework similar to the one implemented for the Unified Enterprise Survey's "Integrated Questionnaire" at Statistics Canada (see Boltwood, 1998) was suggested. Using this framework to plan forms (questionnaires) would result in a well organized and unified, yet completely flexible, approach to business survey data collection. The framework employs modules, where a module is a set of survey content items (questions) on the same topic. A form (questionnaire) is a set of modules, designed for a specific target population, e.g. industry. The forms (questionnaires) can be customized to suit respondent requirements, e.g. complex enterprises may be subject to variable collection requirements.

The enterprises with large number of establishments are more likely to operate in multiple regions and/or multiple industries. In fact, it would suffice to get a consolidated report for the enterprises operating in single region and single industry irrespective of the number of establishments. Thus, the issue is not the enterprises with large number of establishments but the enterprises operating in multiple regions and/or industries (i.e., complex enterprises). For the enterprises consisting of too many establishments it would not be feasible and/or economical to collect data for all establishments. We considered two options for such enterprises: Option 1 was to obtain consolidated reports, and Option 2 was sub-sampling of establishments. Although both options are methodologically sound, Option 2 has a number of drawbacks as discussed below.

1. It is a two-step procedure during field data collection. First the field enumerators will compile the listings of the establishments along with the region and industry codes. These listings will then be sent to head office for selecting a sample of establishments, which will be sent back to the field for data collection which could pose logistical problems.
2. It is less efficient because of additional variance introduced due to sub-sampling of establishments. The computation of variance for sub-sampling establishments belonging to the enterprises in the Must-Take and Take-All strata will be straightforward because the enterprise will become a stratum. But, it will be somewhat complicated when the enterprise was sampled from one of the Take-Some strata.
3. It may be less burdensome for the respondents to provide a consolidated national report for each industry with the corresponding regional breakdown in percentage terms than provide individual establishment's financial reports even if these are for sampled establishments only. Moreover, sampled enterprise may be unwilling or unable to provide data for the assigned establishment.

Option 1 can be implemented uniformly across the board for all enterprises including those that ***are simple*** enterprises with only a few establishments. There will be no loss of information if consolidated reports are obtained from the **simple** enterprises at the national level. Thus, Option 1 was implemented for complex enterprises and the enterprises with large number of establishments and the enterprises with only few establishments were given the choice to provide data for individual establishments or provide a consolidated report with regional break down in percentage terms.

4. Sample Weighting and Estimation

4.1 Sample Weighting

After creating the "clean" data file, weights were constructed for the enterprises that participated in the survey (including the out-of-scope enterprises) so that the responses could be properly expanded to represent the entire population of the enterprises that the sample was selected to represent. The sampling weights were the result of calculations involving several factors, including original selection probabilities to define base weights, adjustment for unknown eligibility, nonresponse and ratio estimation. Moreover, the weights of the respondent enterprises in Take-Some1 strata were further adjusted to represent the contribution for the corresponding Take-None strata.

Adjustment for Unknown Eligibility and Nonresponse

The sampled enterprises can be divided into eligible, ineligible (or out-of-scope) and unknown eligibility. It may not always be possible to determine the eligibility of the sampled enterprises. For example, a sampled enterprise that could not be located could have been closed and hence ineligible (out-of-scope) for the survey OR could have relocated and hence eligible for the survey. Further, the eligible enterprises can be respondents or non-respondents. In order to apply the adjustments for unknown eligibility and nonresponse, the sampled enterprises were grouped into the following four response status categories:

1. **Respondents:** This group consists of all eligible sampled enterprises that provided usable survey data.
2. **Non-respondents:** This group consists of all eligible sampled enterprises that did not provide usable survey data. The information that could be obtained was sufficient to ascertain that the enterprise was eligible for the survey.

3. **Ineligible or Out-of-Scope:** This group consists of all sampled enterprises that were ineligible or out-of-scope for the survey, such as out of business and were therefore ineligible for the survey.
4. **Eligibility Unknown:** This group consists of all sampled enterprises whose eligibility could not be determined. For example, sampled enterprises that could not be located were placed in this category.

The adjustment for unknown eligibility and nonresponse was applied in two steps. In the first step the base weights of the enterprises with unknown eligibility (Category 4) was distributed proportionally over those with known eligibility (Categories 1, 2, and 3). Let w_{hi} denote the base weight of the i^{th} enterprise sampled from stratum h. Then the adjustment for unknown eligibility for stratum h was obtained as:

$$A_h^{(ue)} = \frac{\sum_{i \in R_h} w_{hi} + \sum_{i \in N_h} w_{hi} + \sum_{i \in O_h} w_{hi} + \sum_{i \in U_h} w_{hi}}{\sum_{i \in R_h} w_{hi} + \sum_{i \in N_h} w_{hi} + \sum_{i \in O_h} w_{hi}}, \tag{4.1}$$

where R_h represents enterprises from stratum h that were survey respondents (Category 1), N_h represents non-respondent enterprises from stratum h (Category 2), O_h represents ineligible (or out-of-scope) enterprises from stratum h (Category 3), and U_h represents the sampled enterprises from stratum h whose eligibility could not be determined (Category 4). The adjustment factor $A_h^{(ue)}$ to account for unknown eligibility was applied to the base weights of the respondent enterprises (Category 1), non-respondent enterprises (Category 2), and out-of-scope or ineligible enterprises (Category 3). Thus, the base weight w_{hi}^* adjusted for unknown eligibility for Categories 1, 2 and 3 was computed as:

$$w_{hi}^* = A_h^{(ue)} w_{hi} \tag{4.2}$$

The weights of the enterprises with unknown eligibility (Category 4) was set to zero. It should be noted that the adjustment for unknown eligibility was applied within sampling strata. Since the base weights are constants within sampling strata the sums in equation (4.1) can be replaced by the sample counts in the corresponding categories within strata.

In the second step, the adjustment factor to account for the non-respondent enterprises was calculated as the ratio of the sum of the weights (adjusted for unknown eligibility) for respondent enterprises and non-respondent enterprises to the sum of the weights for the respondent enterprises. Thus, the nonresponse adjustment factor $A_h^{(nr)}$ for stratum h was calculated as:

$$A_h^{(nr)} = \frac{\sum_{i \in R_h} w_{hi}^* + \sum_{i \in N_h} w_{hi}^*}{\sum_{i \in R_h} w_{hi}^*}, \tag{4.3}$$

where w_{hi}^* is the weight obtained after applying the adjustment for unknown eligibility. The adjustment factor $A_h^{(nr)}$ was applied only to the weights of the respondent enterprises (Category 1) in the sample. That is, the nonresponse-adjusted weight w_{hi}^{**} was computed as:

$$w_{hi}^{**} = \begin{cases} A_h^{(nr)} w_{hi}^{*} & if\ i \in R_h \\ w_{hi}^{*} & if\ i \in O_h \end{cases}. \tag{4.4}$$

Thus, the weights of the out-of-scope/ineligible enterprises were the weights obtained after applying the adjustment factor for unknown eligibility. The out-of-scope enterprises were kept on the sample weighted file because these contribute to the variances of the survey estimates. The weights of the non-respondent enterprises were set to zero. In other words, the non-respondent enterprises could be deleted from the sample weighted file.

Weight Adjustment for Ratio Estimates

The base weights adjusted for unknown eligibility and nonresponse were further adjusted to compute separate ratio estimates. The enterprise level total employment counts from the sampling frame was used as auxiliary variable to compute the separate ratio adjustment. We denote by C_h the total employment count in stratum h from the sampling frame. The total employment count C_h can also be estimated from the sample as $\hat{C}_h = \sum_{i \in R_h} w_{hi}^{**} c_{hi} + \sum_{i \in O_h} w_{hi}^{**} c_{hi}$. The adjustment for separate ratio estimation was computed as the ratio:

$$A_h^{(sr)} = C_h \big/ \hat{C}_h \,. \tag{4.5}$$

Thus, the weights adjusted for separate ratio estimation, say w_{hi}^{***} were computed as:

$$w_{hi}^{***} = A_h^{(sr)} w_{hi}^{**}. \tag{4.6}$$

These weights were the final weights that were used for computing survey estimates and the corresponding estimated variances. It should be emphasized that the final weights were constructed both for the respondent and the out-of-scope (ineligible) enterprises. The out-of-scope enterprises do not contribute to the survey estimates but these do contribute to the variances of the estimates. The weights of the non-respondent enterprises were set to zero. In other words, the non-respondent enterprises could be deleted from the weighted sample file.

Take-None Estimation

Since no sample was selected from the Take-None strata, synthetic estimates were constructed for the Take-None strata. The synthetic estimates are based on certain assumptions, and will be subject to synthetic estimation bias if the assumptions did not hold. The bias would be negligible in our application because the contribution from the Take-None strata were kept less than 5.0 percent. We implemented the synthetic estimation by adjusting the weights of the respondent and out-of-scope enterprises within the corresponding Take-Some1 stratum by multiplying with a Take-None factor defined as:

$$Adj_{(Take-None)} = 1 + \frac{C_{(Take-None)}}{C_{(Take-Some1)}}, \tag{4.7}$$

where $C_{(Take-None)}$ and $C_{(Take-Some1)}$ are respectively the total employment counts based on the frame for the Take-None and Take-Some1 strata within the primary stratum. The Take-None weight adjustment was applied separately within each primary stratum. The Take-None weight adjustment can be incorporated along with the ratio adjustment by using in the numerator in equation (4.6) the sum of total employment counts of the Take-Some1 and Take-None strata (instead of Take-Some1) when computing the ratio adjustment factor for the Take-Some1.

4.2 Survey Estimates – Enterprises

All survey estimates were obtained as domain estimates by using an indicator variable ${}_d\delta_{hi}$, where the post-script d denotes the "estimation domain" and the sub-scripts h and i denote the stratum and the enterprise respectively. The estimation domain can be a geographic domain (e.g., a region) or it can be a characteristic domain (e.g., enterprises in certain size category or certain industry). The estimation domain can also be the intersection of two or more geographic and/or characteristics domains. For example, all enterprises located in a particular region and involved in a certain industry. It should be pointed out that domains are defined on the basis of survey responses and NOT on the basis of sampling frame data. The indicator variable ${}_d\delta_{hi}$ is defined as:

$${}_d\delta_{hi} = \begin{cases} 1 & if\ hi \in d \\ 0 & otherwise \end{cases}. \tag{4.8}$$

The advantage of using the indicator variable is that all estimates can be expressed as the "national" level estimates. The indicator variable ${}_d\delta_{hi}$ will automatically exclude those enterprises that are not part of the estimation domain, and we also obtain the correct estimate of the variance. For the sake of simplicity, we denote by w_{hi} instead of w_{hi}^{***} the final survey weights. Moreover, we use n_h to denote the number of responding enterprises (including out-of-scope) from design stratum h.

Estimation of Totals

As mentioned above, all domain estimates were produced as "national" aggregates by using an indicator variable ${}_d\delta_{hi}$. Thus, the survey estimate for domain d for a characteristic y will be given by:

$${}_d\hat{Y} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}\, {}_d\delta_{hi} y_{hi}\ , \tag{4.9}$$

where h denotes the design stratum, and i is the respondent OR out-of-scope enterprise from stratum h, and L is the total number of design strata. The symbols w_{hi} and y_{hi} represent respectively the survey weight and the observed value of the variable (or characteristic) y for the responding or out-of-scope enterprise. The ${}_d\hat{Y}$ can also be written as:

$${}_d\hat{Y} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}\, {}_d y_{hi}\ , \tag{4.10}$$

where ${}_d y_{hi} = {}_d\delta_{hi} y_{hi}$. It should be noted that ${}_d y_{hi}$ will always be zero for the out-of-scope enterprises. Ratio of totals of two variables can be estimated by ratio of the estimated totals of the two variables.

4.3 Survey Estimates – Establishments

A sampled enterprise is a cluster of establishments with one or more establishments belonging to the enterprise. We assume that either all establishments within the sampled enterprises were enumerated or consolidated reports were obtained from the respondent enterprises at the appropriate estimation level, e.g. region by industry, etc. Similar to the enterprise estimates, all establishment estimates will be obtained as domain estimates by using an indicator variable. Let j denote the establishment within the respondent OR out-of-scope enterprise i from stratum h. Then we define an indicator variable ${}_d\delta_{hij}$ for the domain of interest d as:

$$ {}_d\delta_{hij} = \begin{cases} 1 & if\ hij \in d \\ 0 & otherwise \end{cases}. \tag{4.11}$$

As for the enterprise estimates, all domain estimates for the establishments were also produced as "national" aggregates by using the indicator variable ${}_d\delta_{hij}$ defined above. The survey estimate for domain d for a characteristic y is then given by:

$$ {}_d\hat{Y} = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{M_{hi}} w_{hi}\, {}_d\delta_{hij} y_{hij}\ , \tag{4.12}$$

where M_{hi} is the number of establishments in the respondent OR out-of-scope enterprise i in stratum h, L is the number of design strata, w_{hi} is the enterprise weight, and y_{hij} is the observed value of the variable (or characteristic) y for the establishment. It should be noted that the establishment weight is the same as the corresponding parent enterprise weight. The estimated total ${}_d\hat{Y}$ given above in equation (4.12) can also be written as:

$$ {}_d\hat{Y} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi} \left\{ \sum_{j=1}^{M_{hi}} {}_d\delta_{hij}\ {}_d y_{hij} \right\}. \tag{4.13}$$

If we define the derived variable ${}_d y_{hi}^{*}$ as:

$$ {}_d y_{hi}^{*} = \left\{ \sum_{j=1}^{M_{hi}} {}_d\delta_{hij}\ {}_d y_{hij} \right\}, \tag{4.14}$$

and substitute ${}_d y_{hi}^{*}$ from equation (4.14) into equation (4.13) then the survey estimate ${}_d\hat{Y}$ for the establishment domain d can be written as:

$$ {}_d\hat{Y} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}\, {}_d y_{hi}^{*}\ . \tag{4.15}$$

We notice that the survey estimate for the total of establishment domain given in (4.15) is identical to the one for the enterprise domain given in (4.10) except that ${}_d y_{hi}$ in (4.10) has been replaced by ${}_d y_{hi}^{*}$ in (4.15). Thus, the survey estimates for the totals of establishment domains can be written as survey estimates for the totals of enterprise domains by defining the appropriate derived variables at the enterprise level. Once the derived variables have been defined for estimating the totals for the establishment domains the methodology for the estimation of ratios is the same as that employed for the estimation of ratios for the enterprise domains.

4.4 Variance Estimation

The estimated variances of the survey estimates (both enterprise and establishment) were computed using the commercially available SUDAAN[1] statistical software. The

[1] Designed and developed at RTI International, SUDAAN® is an internationally recognized statistical software package that specializes in providing efficient and accurate analysis of data from complex studies.

SUDAAN software employs Taylor Series approximations (or Taylor linearization) by using what are referred to as weighted survey totals. The general approach used by SUDAAN is to form weighted totals (synthetic replicate scores) of the form:

$$w_{hi} z_{hi}^{\left({}_d\theta\right)}, \tag{4.16}$$

where $z_{hi}^{\left({}_d\theta\right)}$ is the value of a synthetic variable (defined later in the current subsection) for the i^{th} enterprise in the design stratum h, and w_{hi} is the corresponding survey weight. The super-script ${}_d\theta$ used with the synthetic variable $z_{hi}^{\left({}_d\theta\right)}$ indicates that its definition depends on the domain parameter to be estimated.

The variances of the survey estimates for establishments can also be obtained in exactly the same manner by Taylor linearization except that we will have to define synthetic variables for the corresponding enterprise level derived variables that were used to construct the establishment level estimates. The SUDAAN statistical software can then be used to construct survey estimates and compute variances of the estimates.

Synthetic Scores for Variance Estimation of Estimated Totals

As mentioned above, all domain estimates were produced as "national" aggregates by using an indicator variable ${}_d\delta_{hi}$. The survey estimate of total for domain d for a characteristic y is given in equation (4.11). It should be emphasized that the ${}_d y_{hi}$ value is always zero for the out-of-scope enterprises. Then the corresponding synthetic variable $z_{hi}^{\left({}_dY\right)}$ for variance estimation using Taylor method is defined as:

$$z_{hi}^{\left({}_dY\right)} = {}_d y_{hi} - {}_d R_h c_{hi} \tag{4.17}$$

The symbols h and d denote respectively the stratum and the estimation domain. The variable c_{hi} is the auxiliary variable (i.e., total employment) at the enterprise level. The ratio ${}_d R_h$ is unknown, and is substituted by its estimate ${}_d \hat{R}_h$ defined as ${}_d\hat{Y}_h \big/ C_h$. It should be noted that the survey estimate of C_h is also equal to C_h due to ratio estimation. The variance of the estimated total ${}_d\hat{Y}$ (assuming SRS) is then given by:

$$v({}_d\hat{Y}) = \sum_{h=1}^{L} (1 - f_h) \times \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(w_{hi} z_{hi}^{\left({}_dY\right)} - \frac{1}{n_h} \sum_{i=1}^{n_h} w_{hi} z_{hi}^{\left({}_dY\right)} \right)^2, \tag{4.18}$$

where f_h is the sampling fraction for stratum h, and the factor $(1 - f_h)$ is known as the finite population correction.

It should be noted that n_h is the number of responding and ineligible (out-of-scope) units from stratum h and not the number initially sampled. It should also be noted that the records corresponding to the sampled units that were observed to be out-of-scope must be included in the data file for estimation. These businesses do not contribute to the estimates but these do contribute to the variances of the estimates.

Acknowledgements

The work was done under a contract with the Philippines National Statistical Office. Author is grateful to Professor J.N.K. Rao for reviewing the paper, and providing valuable suggestions.

References

Boltwood, A. (1998). *A Framework for Planning Unified Enterprise Survey Data Collection.* Internal Report, Enterprise Statistics Division, Statistics Canada, Ottawa.

Choudhry, G.H., Nyfjäll, M. and Winglee, M. (2004). Sample Design for the Terrorism Risk Insurance Program Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 3358-3364.

Choudhry, G.H., Rao, J. N. K. and Hidiroglou, M.A. (2012). On Sample Size Allocation for Efficient Domain Estimation. Survey. *Survey Methodology,* **38**, 23-29.

Dalenius, T. and Hodges, J. L. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association,* **54**, 88-101.

Horgan, M. (2006). *Stratification of Skewed Population: A Review.* International Statistical Review, **74**, 67-76.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Lavallee, P. and Hidiroglou, M.A. (1988). On the Stratification of Skewed Population. *Survey Methodology*, **14**, 33-43.

Wolter, K. M. (1985). *Introduction to Variance Estimation.* Springer-Verlag: New York.