Revista Colombiana de Estadística
Diciembre 2011, volumen 34, no. 3, pp. 403 a 420

Hierarchical Design-Based Estimation in Stratified Multipurpose Surveys

Estimación jerárquica basada en el diseño muestral para encuestas estratificadas multi-propósito

Hugo Andrés Gutiérrez[a], Hanwen Zhang[b]

Centro de Investigaciones y Estudios Estadísticos (CIEES), Facultad de Estadística, Universidad Santo Tomás, Bogotá, Colombia

Abstract

This paper considers the joint estimation of population totals for different variables of interest in multi-purpose surveys using stratified sampling designs. When the finite population has a hierarchical structure, different methods of unbiased estimation are proposed. Based on Monte Carlo simulations, it is concluded that the proposed approach is better, in terms of relative efficiency, than other suitable methods such as the generalized weight share method.

***Key words*:** Design based inference, Finite population, Hierarchical population, Stratified sampling.

Resumen

Este artículo considera la estimación conjunta de totales poblacionales para distintas variables de interés en encuestas multi-propósito que utilizan diseños de muestreo estratificados. En particular, se proponen distintos métodos de estimación insesgada cuando el contexto del problema induce una población con una estructura jerárquica. Con base en simulaciones de Monte Carlo, se concluye que los métodos de estimación propuestos son mejores, en términos de eficiencia relativa, que otros métodos de estimación indirecta como el recientemente publicado método de ponderación generalizada.

***Palabras clave*:** inferencia basada en el diseño, población finita, población jerárquica, muestreo estratificado.

[a]Lecturer. E-mail: hugogutierrez@usantotomas.edu.co
[b]Lecturer. E-mail: hanwenzhang@usantotomas.edu.co

1. Background

The reality of surveys is complex; as Holmberg (2002) states, most of the real applications in survey sampling involve not one, but several characteristics of study; and as Goldstein (1991) claims, real populations have hierarchical structures. Moreover, in certain occasions, the survey methodologist is faced with the estimation of several parameters of interest in different levels of the population and he/she is commanded with the seeking of proper approaches to estimate those parameters as required in the study. The problem of proposing sampling strategies (optimal sampling design and efficient estimators) that contemplate joint estimation of several parameters in multipurpose survey has been widely discussed in recent statistical literature. Although there is a vast number of papers about estimation of hierarchical populations (Gelman & Hill 2006) and model-based (or model-assisted) multilevel survey data (Skinner, Holt & Smith 1989, Lehtonen & Veijanen 1999, Goldstein 2002, Rabe-Hesketh & Skrondal 2006), the design-based estimation for finite populations with hierarchical structures seems to be omitted by survey statisticians. The aim of this paper is to provide a multipurpose approach to the joint estimation of several parameters for different variables in a stratified finite population with two levels.

Next are detailed some clarifying ideas concerning the concept of hierarchical structures in finite populations. Many kinds of data have a hierarchical or clustered structure. Note that in biological studies it is natural to think in a hierarchy where the offspring of the races is clustered into families; in educational surveys, students belong to schools and schools belong to districts, and so on; in social studies, a person belongs to a household and households are grouped geographically. In this paper, the concept of hierarchy is related with the multipurpose approach in the sense that the survey statistician often needs to make inferences on different levels of the finite population. For example, consider an establishment survey. It would be of interest to estimate the total sales of the market sections of the stores in detail (sales by toys, grocery, electronics or pharmacy sections) and at the same time it would be of interest to estimate the number of employees working in the stores. It is clear that the multipurpose approach is given by the joint inference of two different study variables (sales by market section and number of employees in the stores) but these variables of interest are in different levels of the population: sales are related with the market section level and the number of employees with the store level. Note that as the market sections belong to the stores, then the set of all market sections defines the second level and the set of all stores defines the first level.

In some occasions, it is impossible to obtain a sampling frame for the first level, however this is available for the second level. For example, Särndal, Swensson & Wretman (1992, example 1.5.1) reports on the Swedish household survey where there is not a good complete list of households and the sampling frame used was the Swedish Register of the Total Population, which is a list of individuals. In this case, the first level is composed of households, the second level is composed of individuals and the inferences about households are induced directly from the population of individuals. If the requirements of that survey were to obtain inferences about both

households and individuals, then it would be a clear example of a study involving multipurpose estimation within a hierarchical structure in the finite population, with the restriction that the sampling frame is only available in the second level. In other cases, it is possible that both sampling frames are available in the design stage. However, if the requirements of the survey are focused in the estimation of the population totals in both levels, the most trivial, but in some cases useless, solution would be planning two sampling designs. In this paper we propose another solution requiring just the use of a sampling frame in order to simultaneously estimate several parameters for different study variables in two different levels of a stratified population, when the sampling frame to be used is related with the units of the second level. Note that, since the sampling frame is not available (or available but useless) in the first level, sampling designs such as cluster, or multi-stage sampling designs are no longer valid to solve this kind of problems.

The outline of this paper is as follows: after a brief introduction explaining the hierarchical concept, different levels of estimation in such populations, and its implications in the survey sampling context; Section 2, explains in detail, by means of a simple example, the foundations of the hierarchical finite population and the issue of this paper. Section 3, refers to the proposal of an indirect estimation in the first level involving different variables of interest than those considered in the second level. This approach is based on the computation of the first and second order inclusion probabilities, given by the induced sampling design in the first level, using the principles of the well-known Horvitz-Thompson and Hájek estimators for a population total. Besides, in this section, the authors show how this problem is related with the indirect sampling approach (Lavallée 2007). This section also presents a simple case study to illustrate the procedures of the proposed approach in the case of simple random stratified sampling (STSI) in the second level. In Section 4, we present an empirical study based on several Monte Carlo simulations that show how our proposal outperforms, in the sense of relative efficiency, other methods of indirect estimation such as the generalized weight share method (indirect sampling). Finally, some recommendations and conclusions are given in Section 5.

2. Multipurpose Estimation

Let $U = \{1, \ldots, k, \ldots, N\}$ denote the second level finite population of N elements in which a sampling frame is available. Suppose that the sampling frame is stratified and for each element $k \in U$ the stratum to which k belongs is completely identified by means of some discrete auxiliary variable. That is, the population U is partitioned into H subsets $U_1, U_2, ..., U_H$ called strata, where

$$\bigcup_{h=1}^{H} U_h = U, \qquad U_h \bigcap U_{h'} = \emptyset \quad \text{for all } h \neq h'$$

On the other hand, assume that each element $k \in U$ in the second level belongs to a unique cluster in the first level. It is assumed that there exist N_I clusters

denoted by $U_1, \ldots, U_i, \ldots, U_{N_I}$. This set of clusters is symbolically represented as $U_I = \{1, \ldots, i, \ldots, N_I\}$. This way, the first level population is U_I, the second level population is U and, clearly, the data show a notorious hierarchical structure.

Although there is an available sampling frame for U, suppose that it is impossible to obtain a frame for the population of the first level U_I and that the requirements of the survey imply the inference of parameters, say population totals or means, for both levels. Hence, it is assumed that there are two variables of interest, say, y in the second level, and z in the first level, and it is requested the estimation of both population totals, defined by

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k$$

and

$$t_z = \sum_{i \in U_I} z_i$$

In this paper, the notation of any pair of elements in the second level will be denoted by the letters k and l; meanwhile for the units in the first level, the letters i and j will be used.

By taking advantage of the sampling frame in the second level, a stratified sample s is drawn. For each $k \in s$, the value of the variable of interest y_k is observed. Besides, it is supposed that unit k can also provide the information of its corresponding cluster, say U_i. This way, the value of the other variable of interest z_i is recorded. Note that for a particular second level sample there exists a corresponding set of units in the first level. In other words, the second level sample s induces a set, contained in the first level population, which will be called the first level sample, denoted by m and given by

$$m = \{i \in U_I \mid \text{at least one unit of the cluster } U_i \text{ belong to } s\}$$

In summary, the values of both variables of interest could be recorded ar the same time: y_k for the elements in the selected sample; s and z_i for the clusters in the induced sample m. As an example, consider the finite population showed in Table 1. The second level population, denoted by $U = \{A1, B1, D1, \ldots, D4, E4\}$ of size $N = 15$ is a set of market sections in different stores. This population is stratified in four sections ($H = 4$). The population of the first level is hence $U_I = \{A, B, C, D, E\}$ with $N_I = 5$. Each stratum is present in different clusters. For example, Section 1 is present in four stores, whereas Section 3 is present in three stores. Notice that it is not required that each stratum be present in all of the clusters.

Following with the example, when a sample s is drawn, an interviewer visits the selected market section, say k, records the value of y_k and also obtains the information about z_i, the value of the variable of interest in the cluster that contains that section. Table 2, reports the first and second level population values for the variables of interest. If the sampling design is such that only one element

TABLE 1: Description of a possible hierarchical configuration.

	Section 1	Section 2	Section 3	Section 4
Store A	A1	A2	-	A4
Store B	B1	-	B3	-
Store C	-	C2	-	C4
Store D	D1	D2	D3	D4
Store E	E1	E2	E3	E4

of each section is selected, then a possible sample in the second level would be $s = \{A1, E2, B3, E4\}$. This way, the recorded values for this specific sample correspond to 32, 33, 26, 55 and the induced first level sample would be $m = \{A, B, E\}$ and the values of the variable of interest in this level correspond to 14.12, 10.25 and 24.81, respectively. Note that a store may be selected more than once; however, following Särndal et al. (1992, section 3.8), we omit the repeated information in the first level and carry out the inference by using the reduced sample. The parameter of interest in the first level is $t_z = 14.12 + 10.25 + 17.52 + 22.58 + 24.81 = 89.28$ and the parameter of interest in the second level is $t_y = 106 + 105 + 68 + 162 = 441$.

TABLE 2: Variables of interest in a possible hierarchical configuration.

Y_1	Y_2	Y_3	Y_4	Z
$y_{A1} = 32$	$y_{A2} = 12$	-	$y_{A2} = 51$	$Z_A = 14.12$
$y_{B2} = 18$	-	$y_{B3} = 26$	-	$Z_B = 10.25$
-	$y_{C2} = 36$	-	$y_{C4} = 10$	$Z_C = 17.52$
$y_{D1} = 42$	$y_{D2} = 24$	$y_{D3} = 14$	$y_{D4} = 46$	$Z_D = 22.58$
$y_{E1} = 14$	$y_{E2} = 33$	$y_{E3} = 28$	$y_{E4} = 55$	$Z_E = 24.81$

As stated at the beginning of this section, the second level population U is stratified into H strata. In each stratum h $(h = 1, \ldots, H)$ a sampling design $p_h(\cdot)$ is applied and a sample s_h is drawn. An important feature of stratified sampling design is the independence between selections. For this reason, the sampling design takes the following form

$$p(s) = \prod_{h=1}^{H} p_h(s_h) \qquad \text{where} \quad s = \bigcup_{h=1}^{H} s_h$$

We have that an unbiased estimator of t_y and its variance are given by

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} \sum_{s_h} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \hat{t}_{h\pi} \tag{1}$$

$$V(\hat{t}_{y_\pi}) = \sum_{h=1}^{H} V_h(\hat{t}_{h\pi}) = \sum_{h=1}^{H} \sum_{k \in U_h} \sum_{l \in U_h} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$, and $\hat{t}_{h\pi}$ corresponds to the Horvitz-Thompson estimator in the h-th stratum, defined by

$$\hat{t}_{h\pi} = \sum_{s_h} \frac{y_k}{\pi_k}$$

In the case that the sample design is simple random sampling carried out along the strata, the first and second order inclusion probabilities are given by

$$\pi_k = P(k \in s) = P(k \in s_h) = \frac{n_h}{N_h}$$

And

$$\pi_{kl} = \begin{cases} \frac{n_h}{N_h} & \text{if } k = l \\ \frac{n_h}{N_h}\frac{n_h-1}{N_h-1} & \text{if } k \neq l, \text{ with } k, l \in h \\ \frac{n_h}{N_h}\frac{n_{h'}}{N_{h'}} & \text{if } k \neq l, \text{ with } k \in h \text{ y } l \in h' \end{cases}$$

where N_h and n_h denote the population size and the sample size in the stratum h, respectively.

3. Estimation in the First Level

In this section, we develop the proposed approach in order to estimate the parameter of interest in the first level and we point out that another suitable approach could be used to solve this kind of estimation problems, namely the Generalized Weight Share Method (GWSM) (Deville & Lavallée 2006). However, as it will be confirmed later, in the simulation report of Section 4, our proposal is more efficient than the GWSM.

3.1. Proposed Approach

Recalling that the second level sample s induces a first level sample m, we can obtain the induced sampling design as stated in the following result.

Result 1. *The sampling design in the first level induced by the stratified sample s is given by*

$$p(m) = \sum_{\{s:\ s \to m\}} \prod_{h=1}^{H} p_h(s_h) \tag{2}$$

where the notation $s \to m$ indicates that the second level sample s induces the first level sample m.

Proof. Considering that even though a particular first level sample m may be induced by different samples in the second level, it is clear that a second level

sample s may only induce a unique first level sample m, then we have that

$$\begin{aligned} p(m) &= \sum_{\{s:\ s\to m\}} p(s) \\ &= \sum_{\{s:\ s\to m\}} \prod_{h=1}^{H} p_h(s_h) \end{aligned}$$

The last equation follows because of the independence in the selection of s_h for $h = 1, \ldots, H$. □

For example, continuing with the population described in Table 1, if the sampling design in the second level is simple random sampling in each stratum such that $N_3 = 3$, $N_1 = N_2 = N_4 = 4$ and $n_h = 1$ for $h = 1, 2, 3, 4$, then in order to compute the selection probability of the particular first level sample $m = \{A, B\}$, it is necessary to find all of the second level samples inducing that specific sample m. Given the data structure, the set $\{s :\ s \to m\}$ has only two second level samples; these samples are: $\{A1, A2, B3, A4\}$ and $\{B1, A2, B3, A4\}$. For that m, we have that its selection probability corresponds to

$$\begin{aligned} p(m) &= p(\{A1, A2, B3, A4\}) + p(\{B1, A2, B3, A4\}) \\ &= \prod_{h=1}^{4} \frac{1}{N_h} + \prod_{h=1}^{4} \frac{1}{N_h} = \frac{1}{96} = 0.0104 \end{aligned}$$

Given that one parameter of interest is the population total of the variable z in the first level, we can obtain the first and second order inclusion probability of clusters in U_I in order to propose some estimators for t_z. These inclusion probabilities are given in the following results.

Result 2. *The first order inclusion probability of the cluster U_i, denoted by π_i, is given by*

$$\pi_i = Pr(i \in m) = 1 - \prod_{h=1}^{H} q_h^{(i)} \tag{3}$$

where $q_h^{(i)} = Pr(\text{None of the units of } U_i \text{ belongs to } s_h)$ and s_h denotes the selected sample in the stratum U_h, for $h = 1, \ldots, H$.

Proof.

$$\begin{aligned} \pi_i = Pr(i \in m) &= Pr(\text{At least one unit of } U_i \text{ belongs to } s) \\ &= 1 - Pr(\text{None of the units of } U_i \text{ belongs to } s) \\ &= 1 - \prod_{h=1}^{H} q_h^{(i)} \end{aligned}$$

□

***Note* 1.** Note that the computation of the quantities $q_h^{(i)}$ depends on the sampling design used in each stratum. Moreover, if $a_h^{(i)}$ denotes the number of units of cluster U_i belonging to stratum U_h, then $a_h^{(i)} \geq 0$. Which implies that each cluster is not necessarily present in each stratum.

***Note* 2.** The stratified sampling design on the second level population implies independence across strata. However, depending on the sampling design used within each stratum, the independence of units selection may not be guaranteed. For example, in the case of simple random sampling designs, there is no independence. On the other hand, other sampling designs such as Bernoulli and Poisson do provide that independence feature.

Result 3. *The second order inclusion probability for any pair of clusters U_i, U_j is given by*

$$\pi_{ij} = 1 - \prod_{h=1}^{H} q_h^{(i)} - \prod_{h=1}^{H} q_h^{(j)} + \prod_{h=1}^{H} q_h^{(ij)} \tag{4}$$

With $q_h^{(ij)} = Pr(\text{None of the units of } U_i \text{ belongs to } s_h \text{ and none of the units of } U_j \text{ belongs to } s_h)$ and $q_h^{(i)}$, $q_h^{(j)}$ are defined analogously in Result 3.2.

***Proof*.** After some algebra, we have that

$$\begin{aligned}
\pi_{ij} &= Pr(i \in m, j \in m) \\
&= 1 - Pr(i \notin m \text{ or } j \notin m) \\
&= 1 - [Pr(i \notin m) + Pr(j \notin m) - Pr(i \notin m, j \notin m)] \\
&= 1 - [(1 - \pi_i) + (1 - \pi_j) - Pr(i \notin m, j \notin m)] \\
&= 1 - \prod_{h=1}^{H} q_h^{(i)} - \prod_{h=1}^{H} q_h^{(j)} + Pr(i \notin m, j \notin m) \\
&= 1 - \prod_{h=1}^{H} q_h^{(i)} - \prod_{h=1}^{H} q_h^{(j)} + \prod_{h=1}^{H} q_h^{(ij)}
\end{aligned}$$

□

Once these inclusion probabilities are computed, it is possible to estimate t_z by means of the well known Horvitz-Thompson estimator given by

$$\hat{t}_{z\pi} = \sum_{i \in m} \frac{z_i}{\pi_i} \tag{5}$$

Note that $\hat{t}_{z\pi}$ is unbiased for t_z and, if the stratified sampling design in the second level is such that $n_h \geq 2$ for $h = 1, \ldots, H$, its variance is given by

$$V(\hat{t}_{z_\pi}) = \sum_{i \in U_I} \sum_{j \in U_I} \Delta_{ij} \frac{z_i}{\pi_i} \frac{z_j}{\pi_j}$$

Where $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$. However, since the first level sample is induced by the second level sample, the size of m is random, even when the stratified sample design of the second level is of fixed size. For a more detailed discussion about the randomness of the sample size and its effects when a Horvitz-Thompson estimator is used, an interested reader can see Särndal et al. (1992, Example 5.7.3 and Example 7.4.1). In order to avoid extreme estimates, sometimes obtained with the previous estimator, and taking into account that N_I is known, we propose to use the expanded sample mean estimator (denoted in this paper as Hájek estimator) given by

$$\widetilde{t}_z = N_I \frac{\widehat{t}_{z\pi}}{\widehat{N}_{I,\pi}} \tag{6}$$

Where $\widehat{N}_{I,\pi} = \sum_{i\in m} \frac{1}{\pi_i}$. It is well known that its approximate variance is given by

$$AV(\widetilde{t}_z) = \sum_{i\in U_I}\sum_{j\in U_I} \Delta_{ij} \frac{z_i - \overline{z}_{U_I}}{\pi_i} \frac{z_j - \overline{z}_{U_I}}{\pi_j} \tag{7}$$

With $\overline{z}_{i\in\ U_I} = \sum_{U_I} z_i/N_I$. For more comprehensive details, see Gutiérrez (2009, expressions 9.3.7. and 9.3.9.) and Särndal et al. (1992, expression 7.2.10.).

3.1.1. Some Particular Cases

In the case that in each stratum of the second level population a Bernoulli sampling design is used, with the same inclusion probability θ across the strata, then the first order inclusion probability for a cluster U_i is given by

$$\begin{aligned}\pi_i = 1 - \prod_{h=1}^{H} q_h^{(i)} &= 1 - \prod_{h=1}^{H} (1-\theta)^{a_h^{(i)}} \\ &= 1 - (1-\theta)^{\sum_{h=1}^{H} a_h^{(i)}} = 1 - (1-\theta)^{N_i}\end{aligned}$$

Where $N_i = \#(U_i)$. The second order inclusion probability for clusters U_i and U_j is given by

$$\begin{aligned}\pi_{ij} &= 1 - \prod_{h=1}^{H} q_h^{(i)} - \prod_{h=1}^{H} q_h^{(j)} + \prod_{h=1}^{H} q_h^{(ij)} \\ &= 1 - (1-\theta)^{N_i} - (1-\theta)^{N_j} + \prod_{h=1}^{H} (1-\theta)^{a_h^{(i)} + a_h^{(j)}} \\ &= 1 - (1-\theta)^{N_i} - (1-\theta)^{N_j} + (1-\theta)^{N_i+N_j}\end{aligned}$$

Other interesting case is carrying out simple random sampling in each stratum. This way, the resulting formulaes for the proposed approach are quite simple. Denoting the population size and the sample size in the h-th stratum by N_h and

n_h, respectively, and by following the assumptions of the Result 3.2, the first inclusion probability for a cluster U_i is given in terms of $q_h^{(i)}$, where

$$q_h^{(i)} = \begin{cases} \frac{\binom{N_h - a_h^{(i)}}{n_h}}{\binom{N_h}{n_h}}, & \text{if } n_h \leq N_h - a_h^{(i)} \\ 0, & \text{otherwise} \end{cases}$$

On the other hand, for the computation of the second order inclusion probability for clusters U_i and U_j, we have that

$$q_h^{(ij)} = \begin{cases} \frac{\binom{N_h - a_h^{(i)} - a_h^{(j)}}{n_h}}{\binom{N_h}{n_h}}, & \text{if } n_h \leq N_h - a_h^{(i)} - a_h^{(j)} \\ 0, & \text{otherwise} \end{cases}$$

For example, following the finite population in Table 1, the first inclusion probabilities of the store A and store B are given by

$$\pi_{store_{(A)}} = 1 - \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_2}{N_2}\right)\left(1 - \frac{n_4}{N_4}\right)$$
$$\pi_{store_{(B)}} = 1 - \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_3}{N_3}\right)$$

And the second order inclusion probability for these two stores is given by

$$\begin{aligned} \pi_{store_{(A)},store_{(B)}} = 1 &- \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_2}{N_2}\right)\left(1 - \frac{n_4}{N_4}\right) - \left(1 - \frac{n_1}{N_1}\right)\left(1 - \frac{n_3}{N_3}\right) \\ &+ \frac{(N_1 - n_1)}{N_1}\frac{(N_1 - n_1 - 1)}{(N_1 - 1)}\left(1 - \frac{n_2}{N_2}\right)\left(1 - \frac{n_3}{N_3}\right)\left(1 - \frac{n_4}{N_4}\right) \end{aligned}$$

Once the inclusion probabilities are computed, it is possible to obtain estimations of t_z, by using (5) and (6), along with its respective estimated coefficients of variation by means of the expression for the estimated variances.

3.2. Indirect Sampling

This kind of situations can also be handled by using the indirect sampling approach (Lavallée 2007). We introduce it briefly: it is assumed that the first level population U_I is related to the second level population U through a link matrix representing the correspondence between the elements of U_I and U. Since there is no available sampling frame for U_I, an estimate for t_z can be obtained indirectly using a sample from U and the existing links between the two populations. The link matrix is denoted by $\mathbf{\Theta}$ with size $N \times N_I$, and the ki-th element of the matrix $\mathbf{\Theta}$ is defined as

$$[\mathbf{\Theta}]_{ki} = \begin{cases} 1 & \text{if the element } k \text{ is related with the cluster } U_i \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, \ldots, N$, $i = 1, \ldots, N_I$.

The formulation of the standardized link matrix is needed to carry out the estimation of t_z. This matrix is defined as

$$\widetilde{\mathbf{\Theta}} = \mathbf{\Theta}[diag(\mathbf{1}'_N \mathbf{\Theta})]^{-1}$$

where $\mathbf{1}_N$ is the vector of ones of dimension N. It can be shown that $\widetilde{\mathbf{\Theta}}\mathbf{1}_N = \mathbf{1}_{N_I}$. This way, the population total t_z can be expressed as

$$t_z = \mathbf{1}'_{N_I}\mathbf{z} = \mathbf{1}'_N \widetilde{\mathbf{\Theta}}\mathbf{z}$$

Where $\mathbf{z} = (z_1, \ldots, z_{N_I})$. By using the previous expression and taking into account the principles of GWSM, as pointed in Deville & Lavallée (2006), we have the following estimator:

$$\widehat{t}_z = \mathbf{1}'_N \mathbf{I}_N \mathbf{\Pi}_N^{-1} \widetilde{\mathbf{\Theta}}\mathbf{z} \tag{8}$$

where $\mathbf{\Pi}_N = diag(\pi_1, \ldots, \pi_N)$, is a matrix of dimension $N \times N$ that contains the inclusion probabilities for all the elements in the second level population and $\mathbf{I}_N$ is the diagonal matrix containing the indicator variables I_k for the membership of elements in the second level sample s. Note that (8) may be expressed as

$$\widehat{t}_z = \mathbf{w}\mathbf{z}$$

where $\mathbf{w} = \mathbf{1}'_N \mathbf{I}_N \mathbf{\Pi}_N^{-1} \widetilde{\mathbf{\Theta}}$. We can see that the elements of $\mathbf{w}$ are given by

$$w_i = \begin{cases} \sum_{k \in U} I_k \dfrac{\widetilde{\mathbf{\Theta}}_{ki}}{\pi_k}, & \text{if } i \in m \\ 0, & \text{if } i \notin m \end{cases}$$

for $i = 1, \ldots, N_I$. Note that $\widehat{t}_z$ is a weighted sum upon all units in the induced sample m of U_I.

Deville & Lavallée (2006) have shown that $\widehat{t}_z$ is an unbiased estimator for t_z and its variance is given by

$$V(\widehat{t}_z) = \mathbf{z}' \mathbf{\Delta}_{N_I} \mathbf{z}$$

with $\mathbf{\Delta}_{N_I} = \widetilde{\mathbf{\Theta}}' \mathbf{\Delta}_N \widetilde{\mathbf{\Theta}}$, where the kl-th element of $\mathbf{\Delta}_N$ is given by

$$[\mathbf{\Delta}_N]_{kl} = \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l}$$

for $k, l = 1, \ldots, N$.

It is important to comment that despite the resulting inferences of indirect sampling from the GSWM are defined for the first level population, they are directly induced by the probability measure of the sampling design in the second level $p(s)$. However, the inferences from our proposed approach are given directly by the induced sampling design of the first level $p(m)$.

4. Simulation Study

In this section, by means of Monte Carlo simulations, we compare the performance of the two proposed estimators given by (5) and (6) and the indirect sampling estimator. We simulate several stratified populations with hierarchical structure where all clusters are presented in each stratum, that is, $N_h = N_I$ in all strata. The values of the variables of interest y and z are generated from different gamma distributions. Wu (2003) claims that heavy tail distributions such as the log-normal and the gamma distribution with large scale parameters should not be used to generate sampling observations. For this reason, we use the gamma distribution with small shape and scale parameters.

In each stratum, a simple random sample of equal size n is selected, then the two proposed estimators and the indirect sampling estimator are computed in order to estimate t_z. The process was repeated $G = 1000$ times with $N_I = 20, 50, 100, 400$ clusters, and $H = 5, 5, 10, 50$ for each of these values of N_I. The simulation was programmed in the statistical software R (R Development Core Team 2009) and the source codes are available from the author upon request. In the simulation, the performance of an estimator $\widehat{t}$ of the parameter t was tracked by the Percent Relative Bias (RB), defined by

$$RB(\widehat{t}) = 100\% G^{-1} \sum_{g=1}^{G} \frac{\widehat{t}_g - t}{t}$$

and the Relative Efficiency (RE), that corresponds to the ratio of the Mean Square Error (MSE) of the estimator of the GWSM approach to the Horvitz-Thompson and the Hájek estimators defined as

$$RE(\widehat{t}_{z\pi}) = \frac{MSE(\widehat{t}_z)}{MSE(\widehat{t}_{z\pi})} \qquad \text{and} \qquad RE(\widetilde{t}_z) = \frac{MSE(\widehat{t}_z)}{MSE(\widetilde{t}_z)}$$

respectively. Note that $\widehat{t}_g$ is computed in the g-th simulated sample and the Mean Square Error is given by

$$MSE(\widehat{t}) = G^{-1} \sum_{g=1}^{G} (\widehat{t}_g - t)^2$$

The estimators are considered under a wide range of specifications. The simulation results correspond to the ratio of MSE, since the ratio of bias is in all cases negligible indicating that no estimator takes advantage over others in terms of the RB.

Table 3, reports the simulated ratio of MSE for the proposed estimators with the indirect sampling estimator for $N_I = 20$, $H = 5$ and $n = 1, 5, 10, 15$. It can be seen that the Hájek estimator is always more efficient, even when the sample size is $n = 1$. The gain in efficiency increases with increasing sample size. The Horvitz-Thompson estimator has a quite poor performance.

Table 3: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 5$ strata and $N_I = 20$ clusters.

Sample size per stratum	HT	Hájek
n=1	0,08	1,06
n=5	0,03	1,84
n=10	0,05	5,50
n=15	0,52	73,75

Table 4: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 5$ strata and $N_I = 50$ clusters.

Sample size per stratum	HT	Hájek
n=1	0,12	1,02
n=5	0,03	1,29
n=10	0,02	1,57
n=20	0,02	3,24
n=40	1,06	175,83

Table 5: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 10$ strata and $N_I = 100$ clusters.

Sample size per stratum	HT	Hájek
n=1	0,09	1,03
n=10	0,02	1,83
n=20	0,02	3,64
n=50	0,44	101,47

Table 6: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 50$ strata and $N_I = 40$ clusters.

Sample size per stratum	HT	Hájek
n=1	0,02	1,98
n=5	0,77	110,25
n=10	Inf	Inf
n=20	Inf	Inf

Table 7: MSE ratio of the stratified estimator to indirect sampling (IND), HT and Hájek estimators for $H = 5$ strata and $N_I = 20$ clusters.

Sample size per stratum	IND	HT	Hájek
n=1	4,84	3.45	5.39
n=5	4,92	2.53	9.42
n=10	4,34	4.94	27.08
n=15	5,37	40.88	342.90

In the simulation reported in Table 4, we increased the number of clusters to $N_I = 50$, and the sample size to $n = 40$. We see that the Hájek estimator maintains its advantage over the indirect sampling estimator, and it is particularly large when $n = 40$. On the other hand, the Horvitz-Thompson still performs poorly, although when n is close to N_I it is slightly better. The results reported in the Table 5 with $N_I = 100$ and $H = 10$, are similar to those reported in Table 3.

In Table 6, we set $N_I = 40$ and $H = 50$, that is, there are more strata than first level population clusters. We see that the advantage of the Hájek estimator increases substantially even when $n = 5$. The symbol Inf indicates that the MSE of the Horvitz-Thompson and the Hájek estimator are both close to zero in comparison with the MSE of the indirect sampling estimator; that is, the ratio of MSE is huge.)

In order to visualize the average performance of these three approaches, Figure 1, presents the histogram of the Horvitz-Thompson, Hájek and indirect sampling estimators with $N_I = 20$, $H = 5$, $n = 5$. The vertical dotted line indicates the value of the parameter of interest. We observe that the three estimators are unbiased and the estimations obtained with the Hájek estimator are highly concentrated around the population total, while the Horvitz-Thompson estimator has a larger variance.

An interesting, but less practical, situation arises when the parameter of interest in the second level coincides with the parameter of interest in the first level. That is, if $z_i = \sum_{k \in U_i} y_k$, the variable of interest in the cluster U_i corresponds to the total of the variable y in the cluster U_i. In this case, both population totals are the same ($t_y = t_z$) and they can be estimated by using the four mentioned estimators, namely: the stratified estimator given in (1), the Horvitz-Thompson estimator given in (5), the Hájek estimator given in (6) and the indirect sampling estimator given in (8). Notice that in this case, the Horvitz-Thompson, Hájek and indirect estimators use first level information, whereas the stratified estimator uses second level information. Then, it is interesting to evaluate these estimators and compare them. Figure 2 shows the average performance of the four estimators with $N_I = 20$, $H = 5$, $n = 5$. We conclude, once more, that the Hájek estimator is the most efficient and that the estimator of indirect sampling has an acceptable performance, while the stratified and the Horvitz-Thompson estimators have large variances.

Table 7, reports simulation results when comparing the stratified estimator with respect to the remaining three estimators which use the first level information, in terms of relative efficiency. We can see that estimators using first level information are always more efficient than the classical stratified estimator; on the other hand, for each n, the Hájek estimator is the most efficient when increasing the sample size.

The above simulations involve the case that any cluster contains at most one member per stratum, this way the sample includes at most one member in each cluster. However, since our approach may be extended to the general case where a cluster might contain more than one member in some strata, then a more realistic situation arises when we set $a_h > 1$ in some strata. Table 8, reports the simulated

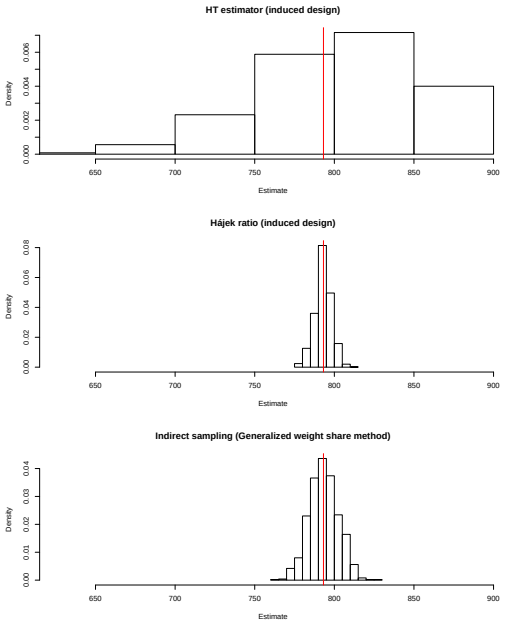

FIGURE 1: Histogram of estimates in 1000 iterations with $N_I = 20$, $H = 5$, $n = 5$.

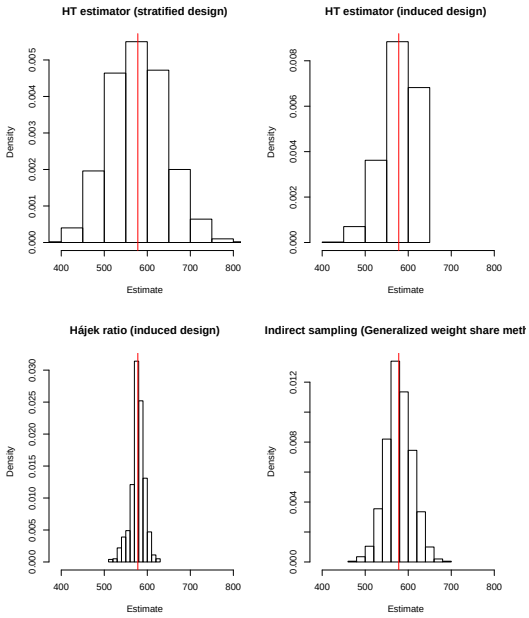

FIGURE 2: Histogram of estimates in 1000 iterations with $N_I = 20$, $H = 5$, $n = 5$.

MSE ratio for the proposed estimators with the indirect sampling estimator for $N_I = 20$, $H = 5$, $a_h = 3$ for each $h = 1, \ldots, H$ and each cluster. Finally, the sample size considered per stratum was $n = 1, 5, 10, 15$. It can be seen that the Hájek estimator is always more efficient, even when sample size is $n = 1$; its gain in efficiency increases with the sample size augmenting. Figure 3, shows the average performance of the three estimators with $N_I = 20$, $H = 5$, $n = 5$.

TABLE 8: MSE ratio of the indirect sampling estimator to HT and Hájek estimators for $H = 5$ strata, $N_I = 20$ clusters and $a_h = 3$.

Sample size per stratum	HT	Hájek
n=1	0,07	1,06
n=5	0,03	1,89
n=10	0,04	4,85
n=15	0,11	17,65

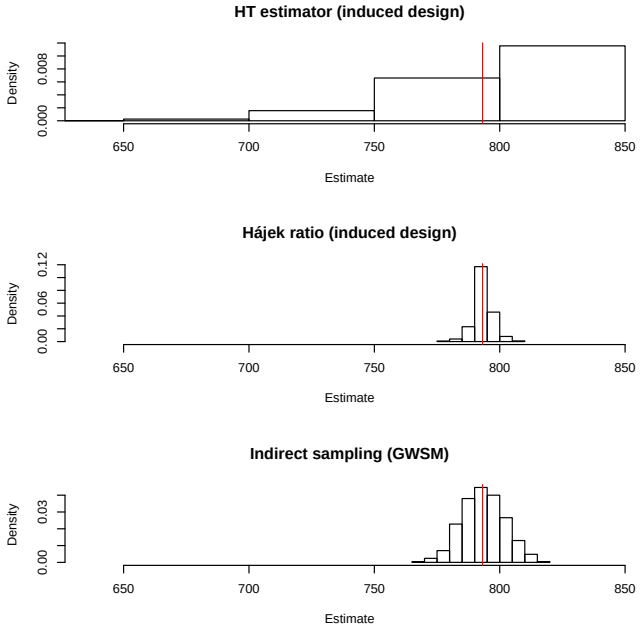

FIGURE 3: Histogram of estimates in 1000 iterations with $N_I = 20$, $H = 5$, $n = 10$ and $a_h = 3$.

It is worth commenting that the Hajek estimator is asymptotically unbiased. However, for samples of size 20 or more, the bias may be important not to be ignored (Särndal et al. 1992, p. 251). There are some proposals available in the literature to modify either the estimator or the sampling design to reduce the bias of this estimator. For a review of some variations of the Hajek estimator, see Rao (1988). Note that even though the sample size in the stratified second

level is small, the induced sample size in the first level is not. This way, it is understandable that the bias for the Hajek estimator is negligible.

5. Discussion and Conclusion

In this paper, we have proposed a design-based approach that yields the unbiased estimation of the population total in the first level based on a stratified sampling design in the second level. With this in mind, the proposed approach is multipurpose in the sense that, for the same survey, different parameters can be estimated in different levels of the population. An important feature of this method is its suitability in the estimation of parameters in the first level where there is no sampling frame available. The empirical study shows that by using the same information, our proposal outperforms the indirect sampling approach because our proposal always has a smaller mean squared error.

The reduction of variability in our proposal may be explained because different second level samples may induce the same first level sample m. In this case, the estimates obtained by applying the GWSM principles will be generally different because the vector of weights $\mathbf{w}$, that depends on the inclusion probabilities of the selected elements in s, differs from sample to sample in the second level. Then we will have different estimates for the same induced sample m. This feature is not present if we follow the approach proposed in this paper, since $\widehat{t}_{z,\pi}$ and $\widetilde{t}_z$ remain constant for different second level samples that induce the same first level sample m. However, $\widehat{t}_{z,\pi}$ does not perform as well as $\widetilde{t}_z$ because, in general, the Horvitz-Thompson approach does not work well under random size sample designs, which is the nature of the sampling design $p(m)$.

This research is still open, further work could be focused in the development of a general methodology conducive to joint estimation in more than two levels when the sampling frame is only available in the last level of the hierarchical population. Besides, the proposed approach could be easily extended in some situations where there is a suitable auxiliary variable (continuous or discrete) that helps to improve the efficiency of the resulting estimators, just as in the functional form of the GWSM with the calibration approach (Lavallée 2007, ch. 7).

Acknowledgements

We thank God for guiding our research. We are grateful to the two anonymous referees for their valuable suggestions and to the Editor in Chief for his advice during the publication process and his comments on the asymptotic unbiased property of the Hajek estimator. Our posthumous gratitude to Leonardo Bautista who motivated this research some years ago. This research was supported by a grant of the Unidad de Investigación from Universidad Santo Tomás.

[Recibido: noviembre de 2009 — Aceptado: mayo de 2011]

References

Deville, J. C. & Lavallée, P. (2006), 'Indirect sampling: the foundation of the generalized weight shared method', *Survey Methodology* **32**(2), 165–176.

Gelman, A. & Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

Goldstein, H. (1991), 'Multilevel modelling of survey data', *Journal of the Royal Statistical Society: Series D (The Statistician)* **40**(2), 235–244.

Goldstein, H. (2002), *Multilevel Statistical Models*, third edn, Wiley.

Gutiérrez, H. A. (2009), *Estrategias de Muestreo. Diseño de Encuestas y Estimación de Parámetros*, Universidad Santo Tomás.

Holmberg, A. (2002), 'A multiparameter perspective on the choice of sampling design in surveys', *Statistics in Transition* **5**, 969–994.

Lavallée, P. (2007), *Indirect Sampling.*, Springer.

Lehtonen, R. & Veijanen, A. (1999), Multilevel-model assisted generalized regression estimators for domain estimation, *in* 'Proceedings of the 52nd ISI Session'.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Rabe-Hesketh, S. & Skrondal, A. (2006), 'Multilevel modelling of complex survey data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**(4), 805–827.

Rao, P. S. R. S. (1988), Ratio and regression estimators, *in* P. R. Krishnaiah & C. Rao, eds, 'Handbook of Statistics', Vol. 6, North-Holland, pp. 449–468.

Särndal, C. E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.

Skinner, C. J., Holt, D. & Smith, T. M. F. (1989), *Analysis of Complex Surveys*, Chichester: Wiley.

Wu, C. (2003), 'Optimal calibration estimators in survey sampling', *Biometrika* **90**(4), 937–951.