Journal of Official Statistics, Vol. 15, No. 4, 1999, pp. 537–550

A Conditional Analysis of Some Small Area Estimators in Two Stage Sampling

Piero D. Falorsi[1] *and Aldo Russo*[2]

The conditional approach in sampling on finite populations analyses the performance of the estimators for the conditional sample space U_c containing samples having some specific properties. The use of conditional arguments in sampling for small area estimation has been studied for the case of simple random sampling. In this article we treat a more realistic situation. We refer to a two-stage sampling design with stratification of the primary sampling units. In each stratum a single primary sampling unit is selected with probability proportional to size. The secondary sampling units are selected without replacement and with equal probabilities. This design is generally used in *household surveys* conducted by the National Statistics Institute of Italy. We consider the following estimators: expansion, ratio, synthetic, and composite expressed as a linear combination of the ratio and synthetic estimators.

The conditional analysis is developed for the reference set U_c that contains all possible samples that have a fixed number of primary sampling units belonging to the small area. We then develop the expressions of the variance and the bias of the four estimators. An empirical analysis concludes the work.

Key words: Conditional bias; conditional variance; conditional mean squared error.

1. Introduction

In the fixed population approach, the sample design defines the sample space U_u (set of all possible samples s) and the associated probabilities of selection $p(s)$.

The evaluation of an estimator $\hat{Y}$ of the total Y is based on the Mean Squared Error (MSE), under repeated sampling with probabilities $p(s)$, using the sample space U_u as the reference set.

Sampling theorists prefer the use of the unconditional approach, based on the unconditional sample space U_u, in planning sampling strategy. However, after the data collection, there is a problem in the choice between the unconditional and conditional approach for the evaluation of the estimator $\hat{Y}$.

The conditional approach is based on the conditional sample space U_c containing samples which have some specific properties.

The use of conditional arguments in sampling has been studied by Holt and Smith (1979) and Royall and Cumberland (1985). The use of the conditional approach for small area estimation has been studied by Rao (1985) and Särndal and Hidiroglou (1989). These

[1] National Statistics Institute, Division on Survey Sampling Methods, Via Balbo 16, 00100 Roma, Italy. E-mail: falorsi@istat.it
[2] Third University of Rome, Via Segre 2, Roma, Italy. E-mail: aldoruss@uniroma3.it
Acknowledgments: Piero D. Falorsi is Senior Researcher and Director of the Division on Survey Sampling Methods of the National Statistics Institute. Aldo Russo is Professor of Statistics at Third University of Rome.

articles consider the case of simple random sampling. In our previous work (Russo and Falorsi 1993) within the context of small area estimation, we studied the conditional and the unconditional properties of some estimators for a simple two stage sampling design without stratification.

The present article is an extension of the previous work of Russo and Falorsi, in which we refer to a two stage sampling design with stratification of the Primary Sampling Units (PSUs). In each stratum a single PSU is selected with probability proportional to size. The Secondary Sampling Units (SSUs) are selected without replacement and with equal probabilities. This kind of design is very important and is generally used in *household surveys* conducted by the National Statistics Institute of Italy. Other relevant surveys based on a multistage sampling design with the selection of a single PSU in the first stage are the *Current Population Survey* of the United States of America and the *Labour Force Survey* of Germany.

We consider the following estimators: expansion, ratio, synthetic, and composite expressed as a linear combination of the ratio and synthetic estimators.

In the sampling context under study it is possible to choose different reference sets. In this work a conditional analysis is developed with respect to a reference set U_c that comprises all possible samples containing a fixed number of PSUs belonging to the small area.

We develop the expressions of the variance and of the bias of the four estimators, which allows us to analyse the conditional theoretical properties of the estimators under study. An empirical analysis concludes the work.

2. Parameter of Interest

We consider a sampling design planned for estimating the total Y_R of an area denoted as R. Our aim is to estimate the total Y_d of a small area, denoted as d, included in R and obtained by an aggregation of PSUs. Each PSU is totally contained within a small area d only. In this context d is an *unplanned domain*, this is an area that was not identified at the time of design and thus may cut across design strata. We denote D as the set of strata including d. (In our notation a symbol denoting a set may also be used for indicating the number of units belonging to the set; the context will clarify the meaning of each symbol.) In order to explain the subsequent algebraic developments, we introduce the following symbols: $h =$ stratum index; $i =$ PSU index; $j =$ SSU index; $N_h =$ set of PSUs of the stratum h; $N_{d,h} =$ set of PSUs of the stratum h belonging to d; $M_{hi} =$ set of SSUs belonging to the PSU hi; $M_h =$ set of SSUs belonging to the N_h PSUs of the stratum h; Y_{hij} value of the variable of interest y in the SSU hij.

Using the above symbols we express Y_d as

$$Y_d = \sum_{h \in D} Y_{d,h} = \sum_{h \in D} \sum_{i \in N_{d,h}} Y_{hi} = \sum_{h \in D} \sum_{i \in N_{d,h}} \sum_{j \in M_{hi}} Y_{hij} \tag{1}$$

3. Conditional Analysis

We denote with s a generic sample selected in D. We denote with $n(n = D)$ the number of sample PSUs of s and with n_d the number of sample PSUs that happen to fall into small area d. The number n_d is a random variable that may assume the values: $0, 1, \ldots, n$.

The conditional analysis is conducted with reference to the conditional sample space U_c containing all samples having a fixed number, say n_d, of PSUs belonging to d.

The conditional probability of drawing the sample s, such that $s \in U_c$, is (Särndal and Hidiroglou 1989, p. 269)

$$p_c(s) = \left[\sum_{s \in U_c} p(s)\right]^{-1} p(s) \tag{2}$$

where $p(s)$ is the unconditional probabilities of drawing the samples s in the sample space U_u.

Therefore, using Expression (2), the conditional inclusion probability of the PSU hi is defined by (Särndal, Swensson, and Wretman 1992, p. 31)

$$\pi_{c,hi} = \sum_{s(hi) \in U_c} p_c(s(hi)) = \left[\sum_{s \in U_c} p(s)\right]^{-1} \sum_{s(hi) \in U_c} p(s(hi)) \tag{3}$$

where $s(hi)$ denotes the generic sample of U_c that contains the PSU hi, $p(s(hi))$ and $p_c(s(hi))$ being respectively the unconditional and the conditional probabilities of drawing the sample $s(hi)$.

In order to derive the expression of the denominator of the right hand term of Formula (3), we observe that it is possible to subdivide U_c into $\binom{n}{n_d}$ subsets of samples.

It is feasible to associate a configuration expressed in terms of strata to each subset. The generic configuration may be represented by means of a sequence of 1 and 0 such as:

Stratum	1	2	...	h	...	n
Generic configuration	1	0	...	1	...	0

where 1 denotes a stratum in which the selected PSU belongs to the small area d and 0 indicates a stratum in which the selected PSU does not fall into the small area. In each configuration there are n_d 1's , and $(n - n_d)$ 0's.

Bearing in mind that a single PSU is selected in each stratum, we denote by δ_h $(h = 1, \ldots, n)$ a dichotomous variable which equals 1 if the selected sample PSU in the stratum h falls into the small area d, and otherwise equals 0. The probability of the generic configuration $g_w(w = 1, \ldots, \binom{n}{n_d})$ is given by

$$p(g_w) = \prod_{h=1}^{n} Z_h^{\delta h}(1 - Z_h)^{1-\delta_h} \tag{4}$$

where

$$Z_h = \sum_{i=1}^{N_{d,h}} \pi_{hi}$$

in which $\pi_{hi} = M_{hi}/M_h$ is the inclusion probability of PSU hi in the unconditional sample space U_u.

Consequently, we have

$$\sum_{s \in U_c} p(s) = \sum_{w=1}^{\binom{n}{n_d}} p(g_w) \tag{5}$$

Example 3.1, following this section, illustrates the use of Expression (5).

We now derive the expression of the factor, $\sum_{s(hi) \in U_c} p(s(hi))$, of Formula (3). First we examine the case in which the PSU hi belongs to the small area d. In this case, given that the PSU hi is selected in the sample, we consider the remaining $(n-1)$ strata. In (n_d-1) of these strata selected sample PSUs fall in small area d; in the remaining $(n-n_d)$ strata selected PSUs do not belong to the small area d. Denote with C_h the subset of the $\binom{n}{n_d}$ configurations defined above, having ''1 in correspondence of the stratum h''; C_h is formed by $(n-1/n_d-1)$ configurations. Let g_v ($v = 1, \ldots, \binom{n-1}{n_d-1}$) denote a generic configuration of C_h and allow g_v(hi) to indicate the set of samples belonging to g_v in which the PSU hi is selected. The sum of the selection probabilities of the samples of the set $g_v(hi)$ is derived from

$$p(g_v(hi)) = \pi_{hi} \prod_{u=1}^{n-1} Z_u^{\delta_u}(1-Zu)^{1-\delta_u} \qquad (u \neq h) \tag{6}$$

Consequently, we have

$$\sum_{s(hi) \in U_c} p(s(hi)) = \sum_{v=1}^{\binom{n-1}{n_d-1}} p(g_v(hi)) \tag{7}$$

Let us now examine the case in which the PSU hi does not belong to the d. In an analogy with the case examined above, we have

$$\sum_{s(hi) \in U_c} p(s(hi)) = \sum_{v=1}^{\binom{n-1}{n_d}} p(g_v(hi)) \tag{8}$$

Indeed, if the sample PSU of the stratum h does not belong to d, this identifies $(n-1)$ strata: in n_d of these strata the sample PSU belongs to d; and in $(n-1-n_d)$ strata the sample PSU does not belong to d.

Example 3.2, following this section, illustrates the use of Expressions (7) and (8).

Example 3.1 Consider the case in which $n = 5$ and $n_d = 3$. The possible $\binom{5}{3} = 10$ configurations are illustrated in the following table.

For example, the number of samples that have the configuration g_1 is given by

$$N_{d,1}N_{d,2}N_{d,3}(N_4 - N_{d,4})(N_5 - N_{d,5}) \tag{9}$$

The sum of quantities similar to (9) give the number of samples belonging to U_c, each of them with size $n = 5$ and $n_d = 3$. The probability of having the configuration g_1 is given by

$$\begin{aligned} p(g_1) &= Z_1 Z_2 Z_3 (1-Z_4)(1-Z_5) \\ &= \left(\sum_{i=1}^{N_{d,1}} \pi_{1i}\right)\left(\sum_{i=1}^{N_{d,2}} \pi_{2i}\right)\left(\sum_{i=1}^{N_{d,3}} \pi_{3i}\right)\left(1-\sum_{i=1}^{N_{d,4}} \pi_{4i}\right)\left(1-\sum_{i=1}^{N_{d,5}} \pi_{5i}\right) \end{aligned} \tag{10}$$

The remaining configurations $(g_2, \ldots, g_{10})$ have expressions similar to (10).

Table 1. Configurations with $n=5$ and $n_{\mathrm{d}}=3$

Configurations	Strata				
	1	2	3	4	5
g_1	1	1	1	0	0
g_2	1	1	0	1	0
g_3	1	1	0	0	1
g_4	1	0	1	1	0
g_5	1	0	1	0	1
g_6	1	0	0	1	1
g_7	0	1	1	1	0
g_8	0	1	1	0	1
g_9	0	1	0	1	1
g_{10}	0	0	1	1	1

Consequently, we have

$$\sum_{s\in U_c} p(s) = \sum_{w=1}^{10} p(g_w) = \sum_{h_1=1}^{3}\sum_{h_2>h_1}^{4}\sum_{h_3>h_2}^{5} Z_{h_1}Z_{h_2}Z_{h_3}(1-Z_{q_1})(1-Z_{q_2})$$

$$= \sum_{h_1=1}^{n-(n_d-1)}\sum_{h_2>h_1}^{n-(n_d-2)}\sum_{h_3>h_2}^{n} Z_{h_1}Z_{h_2}Z_{h_3}(1-Z_{q_1})(1-Z_{q_2}) \tag{11}$$

where h_1, h_2 and h_3 are stratum indexes in which the sample PSU belongs to small area d and q_1, q_2 are stratum indexes in which the sample PSU does not belong to small area d, with $(h_1,h_2,h_3) \neq (q_1,q_2)$ and $(q_1 \neq q_2)$. For example, for the configuration g_5, we have $h_1 = 1$, $h_2 = 3$, $h_3 = 5$, $q_1 = 2$, $q_2 = 4$.

In the general case we have

$$\sum_{s\in U_c} p(s) = \sum_{h_1=1}^{n-(n_d-1)}\sum_{h_2>h_1}^{n-(n_d-2)}\cdots\sum_{h_{n_d}>h_{n_d-1}}^{n} Z_{h_1}Z_{h_2}\ldots Z_{h_{n_d}}(1-Z_{q_1})\ldots(1-Z_{q_{n-n_d}}) \tag{12}$$

with $(h_1,h_2,\ldots.,h_{n_d}) \neq (q_1,\ldots.,q_{n-n_d})$ and $(q_1 \neq q_2 \neq \ldots \neq q_{n-n_d})$

Example 3.2 Consider the case, illustrated in example 3.1, in which $n=5$, $n_d=3$. Consider further the case in which the PSU hi is of the first stratum ($h=1$) and belongs to the small area d; the possible configurations associated with this case are those expressed in Table 1 as $g_1, g_2, g_3, g_4, g_5, g_6$. The probability of having a sample of the configuration g_1, in which the PSU $1i$ belonging to the small area is selected, is given by

$$p(g_1(1i)) = \pi_{1i}Z_2Z_3(1-Z_4)(1-Z_5)$$

$$= \pi_{1i}\left(\sum_{i=1}^{N_{d,2}}\pi_{2i}\right)\left(\sum_{i=1}^{N_{d,3}}\pi_{3i}\right)\left(1-\sum_{i=1}^{N_{d,4}}\pi_{4i}\right)\left(1-\sum_{i=1}^{N_{d,5}}\pi_{5i}\right) \tag{13}$$

For $p(g_2(1i)), p(g_3(1i)), \ldots, p(g_6(1i))$, these expressions are similar to (13), in which the first factor equals π_{1i}. Hence, using expression (7) we have

$$\sum_{s(1i)\in U_c} p(s(1i)) = \sum_{v=1}^{6} p(g_v(1i)) = \pi_{1i} \sum_{h_1=2}^{4} \sum_{h_2>h_1}^{5} Z_{h_1} Z_{h_2} (1-Z_{q_1})(1-Z_{q_2}) \tag{14}$$

with $(h_1, h_2) \neq (q_1, q_2)$ and $(q_1 \neq q_2)$.

For handling the case in which the PSU hi is of the generic stratum ($h = 1, \ldots, 5$) and belongs to the small area d, we denote with: $\breve{h}$ the stratum under study ($\breve{h} = 1$, or $\breve{h} = 2, \ldots, \breve{h} = 5$), we then rearrange the stratum codes giving to generic stratum $h (h \neq \breve{h})$ the code γ expressed by

$$\gamma = \begin{cases} h & \text{for } h < \breve{h} \\ h-1 & \text{for } h > \breve{h} \end{cases} \tag{15}$$

Thus, the probability expressed by (7) is given by

$$\begin{aligned} \sum_{s(1i)\in U_c} p(s(\breve{h}i)) &= \pi_{\breve{h}i} \sum_{\gamma_1=1}^{n-(n_d-2)} \sum_{\gamma_2>\gamma_1}^{n} Z_{\gamma_1} Z_{\gamma_2} (1-Z_{q_1})(1-Z_{q_2}) \\ &= \pi_{\breve{h}i} \sum_{\gamma_1=1}^{4} \sum_{\gamma_2>\gamma_1}^{5} Z_{\gamma_1} Z_{\gamma_2} (1-Z_{q_1})(1-Z_{q_2}) \end{aligned} \tag{16}$$

where (γ_1, γ_2) are stratum indexes, expressed by (15), in which the sample PSU belongs to small area d, and q_2, q_2 are stratum indexes, expressed by (15), in which the sample PSU does not belong to small area d, with $(\gamma_1, \gamma_2) \neq (q_1, q_2)$ and $(q_1 \neq q_2)$.

The formula (16) may be generalised for any n and n_d by

$$\begin{aligned} &\sum_{s(\breve{h}i)\in U_c} p(s(\breve{h}i)) = \\ &= \pi_{\breve{h}i} \sum_{\gamma_1=1}^{n-(n_d-2)} \sum_{\gamma_2>\gamma_1}^{n-(n_d-3)} \cdots \sum_{\gamma_{n_d-1}>\gamma_{n_d-2}}^{n} Z_{\gamma_1} Z_{\gamma_2} \ldots Z_{\gamma_{n_d-1}} (1-Z_{q_1}) \ldots (1-Z_{q_{n-n_d}}) \end{aligned} \tag{17}$$

with the PSU $\breve{h}i$ belonging to small area d, $(\gamma_1, \gamma_2, \ldots, \gamma_{n_d-1}) \neq (q_1, \ldots, q_{n-n_d})$ and $(q_1 \neq q_2 \neq \ldots \neq q_{n-n_d})$.

Adopting a methodology analogous to that described above, it is possible to derive the probability expressed by (8) for the case in which the PSU $\breve{h}i$ does not belong to the small area d; we have

$$\begin{aligned} &\sum_{s(\breve{h}i)\in U_c} p(s(\breve{h}i)) \\ &= \pi_{\breve{h}i} \sum_{\gamma_1=1}^{n-(n_d-1)} \sum_{\gamma_2>\gamma_1}^{n-(n_d-2)} \cdots \sum_{\gamma_{nd}>\gamma_{n_d-1}}^{n} Z_{\gamma_1} Z_{\gamma_2} \ldots Z_{\gamma_{n_d}} (1-Z_{q_1}) \ldots (1-Z_{q_{n-(n_d-1)}}) \end{aligned} \tag{18}$$

with $(\gamma_1, \gamma_2, \ldots, \gamma_{n_d}) \neq (q_1, \ldots, q_{n-(n_d-1)})$ and $(q_1 \neq q_2 \neq \ldots \neq q_{n-(n_d-1)})$

4. Estimators Under Study

We consider the following estimators: expansion (E), ratio (R), synthetic (S) and composite

(C), formally expressed by

$$\hat{Y}_{d,E} = \sum_{h \in D} \sum_{j \in m_{hi}} Y_{hij} \delta_{hi} (\pi_{j,hi} \pi_{hi})^{-1} \tag{19}$$

$$\hat{Y}_{d,R} = (\hat{Y}_{d,E} / \hat{X}_{d,E}) X_d \tag{20}$$

$$\hat{Y}_{d,S} = (\hat{Y}_E / \hat{X}_E) X_d \tag{21}$$

$$\hat{Y}_{d,C} = \alpha \hat{Y}_{d,R} + (1 - \alpha) \hat{Y}_{d,S} \tag{22}$$

being

$$\hat{X}_{d,E} = \sum_{h \in D} \sum_{j \in m_{hi}} X_{hij} \delta_{hi} (\pi_{j,hi} \pi_{hi})^{-1}, \quad \hat{Y}_E = \sum_{h \in D} \sum_{j \in m_{hi}} Y_{hij} (\pi_{j,hi} \pi_{hi})^{-1}$$

$$\hat{X}_E = \sum_{h \in D} \sum_{j \in m_{hi}} X_{hij} (\pi_{j,hi} \pi_{hi})^{-1}, \quad X_d = \sum_{h \in D} \sum_{i \in_d N_h} \sum_{j \in M_{hi}} X_{hij}$$

where: $\pi_{j,hi} = m_{hi}/M_{hi}$ is the inclusion probability of the SSU *hij* conditional on the selection of the PSU *hi*; δ_{hi} is a dichotomous variable that is equal to 1 if the sample PSU belongs to d, otherwise it is equal to 0; X_{hij} is the value of the auxiliary variable x of the SSU *hij*; X_d is the known total of x in d; α is a constant ($0 \leq \alpha \leq 1$). Overviews of options in the choice of α are given by Schaible (1978), Ghosh and Rao (1994) and Singh, Gambino, and Mantel (1994). There are a number of possible approaches in the choice of α. It may be fixed in advance, it may be sample size dependent, or it may be data dependent; the latter two options adapt the estimator to the amount of information available in the sample, so that the ratio estimator is used when it is reliable, and otherwise more weight is given to the synthetic component.

Further, we observe that the symbols $\hat{Y}_E$ and $\hat{X}_E$ denote the expansion estimators of the totals

$$Y_D = \sum_{h \in D} \sum_{i \in N_h} \sum_{j \in M_{hi}} Y_{hij}, \quad X_D = \sum_{h \in D} \sum_{i \in N_h} \sum_{j \in M_{hi}} X_{hij}$$

referred to the area D, formed by the set of strata including the small area d.

5. Conditional Bias

In order to obtain the conditional bias of estimator E, we express this estimator as

$$\hat{Y}_{d,E} = \sum_{h \in D} \sum_{i \in N_{d,h}} (1/\pi_{hi}) \lambda_{hi} \sum_{j \in M_{hi}} (1/\pi_{j,hi}) Y_{hij} \lambda_{hij} \tag{23}$$

where $\lambda_{hi} = 1$, if the PSU is included in the sample and otherwise equals 0; $\lambda_{hij} = 1$, if the SSU hij is included in the sample and otherwise equals 0. The conditional expected value of (23) is given by

$$\begin{aligned} E_c(\hat{Y}_{d,E}) &= \sum_{h \in D} \sum_{i \in N_{d,h}} (1/\pi_{hi}) E_{1c}(\lambda_{hi}) \sum_{j \in M_{hi}} (M_{hi}/m_{hi}) Y_{hij} E_2(\lambda_{hij}) \\ &= \sum_{h \in D} \sum_{i \in N_{d,h}} (\pi_{c,hi}/\pi_{hi}) Y_{hi} \end{aligned} \tag{24}$$

where: E_c denotes averaging over U_c; E_{1c} denotes the conditional expectation over first stage selections; E_2 indicates averaging over second stage selection. Consequently, the conditional bias of the expansion estimator is expressed by

$$B_c(\hat{Y}_{d,E}) = E_c(\hat{Y}_{d,E}) - Y_d = \sum_{h \in D} \sum_{i \in N_{d,h}} (\pi_{c,hi}/\pi_{hi}) Y_{hi} - Y_d \tag{25}$$

In order to obtain the conditional bias of $\hat{Y}_{d,R}$, we consider the linear approximation (Wolter 1985) of the estimator where the partial derivatives are calculated at the conditional mean values. We have

$$\begin{aligned}\hat{Y}_{d,R} = {} & \left[E_c(\hat{Y}_{d,E})/E_c(\hat{X}_{d,E})\right]X_d + \left[X_d/E_c(\hat{X}_{d,E})\right]\left[\hat{Y}_{d,E} - E_c(\hat{Y}_{d,E})\right] + \\ & - X_d\left[E_c(\hat{Y}_{d,E})/E_c^2(\hat{X}_{d,E})\right]\left[\hat{X}_{d,E} - E_c(\hat{X}_{d,E})\right]\end{aligned} \tag{26}$$

The conditional expectation of $\hat{X}_{d,E}$ may be obtained by expression (24) substituting X_{hij} to Y_{hij} and X_{hi} to Y_{hi}. Thus, the conditional expectation of (26) is given by

$$E_c(\hat{Y}_{d,R}) = \left[E_c(\hat{Y}_{d,E})/E_c(\hat{X}_{d,E})\right]X_d \tag{27}$$

Therefore, the conditional bias of the ratio estimator is

$$B_c(\hat{Y}_{d,R}) = \left[E_c(\hat{Y}_{d,E})/E_c(\hat{X}_{d,E})\right]X_d - Y_d \tag{28}$$

Using the linearization method, we can define the conditional bias of the synthetic estimator by

$$B_c(\hat{Y}_{d,S}) = \left[E_c(\hat{Y}_E)/E_c(\hat{X}_E)\right]X_d - Y_d \tag{29}$$

where

$$E_c(\hat{X}_E) = \sum_{h \in D} \sum_{i \in N_h} (\pi_{c,hi}/\pi_{hi}) X_{hi}$$

$$E_c(\hat{Y}_E) = \sum_{h \in D} \sum_{i \in N_h} (\pi_{c,hi}/\pi_{hi}) Y_{hi}$$

Hence the conditional bias of the composite estimator is given by

$$B_c(\hat{Y}_{d,C}) = \alpha B_c(\hat{Y}_{d,R}) + (1-\alpha) B_c(\hat{Y}_{d,S}) \tag{30}$$

6. Variance

To obtain the variance of the expansion estimator, we start with the following expression (Cochran 1977, p. 301)

$$V_c(\hat{Y}_{d,E}) = V_{1c}\left[E_2(\hat{Y}_{d,E})\right] + E_{1c}\left[V_2(\hat{Y}_{d,E})\right]$$

where V_{1c} denotes the conditional first stage variance in U_c, and V_2 denotes the second stage variance for a given set of selected PSUs.

Using the above and the theorem 11.1 cited in Cochran (1977), we obtain

$$V_c(\hat{Y}_{d,E})$$

$$= \sum_{h \in D} \left[\sum_{i \in N_{d,h}} (Y_{hi}/\pi_{hi})^2 \pi_{c,hi}(1-\pi_{c,hi}) - 2 \sum_{i' > i} Y_{hi} Y_{hi'} \pi_{c,hi} \pi_{c,hi'} (\pi_{hi}\pi_{hi'})^{-1} \right.$$

$$+\sum_{i\in N_{d,h}}\left(\pi_{c,hi}/\pi_{hi}^2\right)(M_{hi}-m_{hi})[\pi_{j,hi}(M_{hi}-1)]^{-1}\sum_{j\in M_{hi}}(Y_{hij}-(Y_{hi}/M_{hi}))^2$$

$$\left.+2\sum_{h'>h}\sum_{i\in N_{d,h}}\sum_{i'\in N_{d,h'}}(\pi_{c,hi,h'i'}-\pi_{c,hi}\pi_{c,h'i'})Y_{hi}Y_{h'i'}(\pi_{hi}\pi_{h'i'})^{-1}\right] \tag{31}$$

in which

$$\pi_{c,hi,h'i'}=\sum_{s(hi,h'i')\in U_c}p_c(s(hi,h'i'))=\left(\sum_{s\in U_c}p(s)\right)^{-1}\sum_{s(hi,h'i')\in U_c}(p(s(hi,h'i'))$$

denotes the conditional second order inclusion probabilities of the primary units hi and $h'i'$, being $s(hi,h'i')$ the generic sample of U_c that contains these PSUs and $p(s(hi,h'i'))$ the unconditional selection probability of the sample $s(hi,h'i')$.

Still using the linear approximation, the conditional variance of the ratio estimator may be obtained from (31) by substituting Y_{hij} and Y_{hi} respectively with Z_{hij} and Z_{hi}, expressed by

$$Z_{hij}=\left[X_d/E_c(\hat{X}_{d,E})\right]\left[Y_{hij}-(E_c(\hat{Y}_{d,E})/E_c(\hat{X}_{d,E}))X_{hij}\right]$$
$$Z_{hi}=\sum_{j\in M_{hi}}Z_{hij}$$

The conditional variance of the synthetic estimator is given by

$$V_c(\hat{Y}_{d,S})$$
$$=\sum_{h\in D}\left[\sum_{i\in N_h}(Q_{hi}/\pi_{hi})^2\pi_{c,hi}(1-\pi_{c,hi})-2\sum_{i'>i}Q_{hi}Q_{hi'}\pi_{c,hi'}(\pi_{hi}\pi_{hi'})^{-1}\right.$$
$$+\sum_{i\in N_h}\left(\pi_{c,hi}/\pi_{hi}^2\right)(M_{hi}-m_{hi})(\pi_{j,hi}(M_{hi}-1))^{-1}\sum_{j\in M_{hi}}(Q_{hij}-(Q_{hi}/M_{hi}))^2$$
$$\left.+2\sum_{h'>h}\sum_{i\in N_h}\sum_{i'\in N_{h'}}(\pi_{c,hi,h'i'}-\pi_{c,hi}\pi_{c,h'i'})Q_{hi}Q_{h'i'}(\pi_{hi}\pi_{h'i'})^{-1}\right] \tag{32}$$

where

$$Q_{hij}=\left[X_d/E_c(\hat{X}_E)\right]\left[Y_{hij}-(E_c(\hat{Y}_E)/E_c(\hat{X}_E))X_{hij}\right]$$
$$Q_{hi}=\sum_{j\in M_{hi}}Q_{hij}$$

As far as the variance of the composite estimator is concerned, it may be obtained from (32) by substituting Q_{hij} and Q_{hi} respectively with, W_{hij} and W_{hi}, being

$$W_{hij}=\alpha\,\delta_{hi}Z_{hij}+(1-\alpha)Q_{hij}$$
$$W_{hi}=\sum_{j\in M_{hi}}W_{hij}$$

7. Empirical Study

The evaluation of the conditional performance measures, presented below, of the proposed estimators is carried out for a stratified cluster sample design with strata and cluster delineations and sample sizes identical to those adopted in the 1993 *Multipurpose Household Survey* conducted by the National Statistics Institute of Italy.

This design is based on a two-stage selection with stratification of PSUs. The PSUs are municipalities, while the SSUs are households. In Italy, each geographical region is comprised of municipalities. In every region, the PSUs are divided into two main area types: the *Self-Representing Area* (SRA) consisting of the larger PSUs, and the *Non Self-Representing Area* (NSRA) consisting of the smaller PSUs. All the PSUs in the SRA are sampled, while the selection of the PSUs in the NSRA is carried out within strata that are approximately equal in size. In each stratum only one PSU is selected with probability proportional to size. The SSUs are selected without replacement and with equal probabilities. All members of the selected households are included in the sample.

For the empirical study, the information referring to the sample design, the auxiliary variable x and the variable of interest y are taken from the 1991 General Population Census of Italy.

In our study we consider the region Tuscany as area R, and as small areas the nine provinces of this Region. Because of space constraint, we limit ourselves here to an illustration of these results involving two selected provinces: Florence and Siena. The variable of interest y is the number of people unemployed, and the quantity X_{hij} represents the number of members of the j household in the i municipality belonging to h stratum. The number of strata in the region Tuscany is 50 (consequently we have 50 selected PSUs in the sample); the total number of sample SSUs is equal to 1,452.

We observe that the number of strata in the set D containing the province of Florence is equal to 22; 11 of these strata are entirely comprised of PSUs belonging to the province of Florence; the remaining 11 strata contain both PSUs of this province and PSUs that do not belong to it. Thus, for the province of Florence the number n_d varies in the range 11–22. For the province of Siena, the number of strata in the set D is equal to 12, of which four are entirely composed of PSUs of this province.

For each n_d value we have calculated, by means of a suitable SAS software, all the possible configurations as described in Section 3; consequently, we have obtained the conditional inclusion probabilities $\pi_{c,hi}$ values.

Thus, using the census quantities Y_{hij}, X_{hij}, M_{hi}, $\pi_{hi} = M_{hi}/M_h$, Y_{hi} and X_{hi} and probabilities $\pi_{c,hi}$ we have calculated, for each n_d, the following conditional performance measures:

(i) *Relative Conditional Bias*, defined as

$$\mathrm{RCB}\,(\hat{Y}_{d,m}) = B_c(\hat{Y}_{d,m})/Y_d$$

(ii) *Conditional Standard Error*, expressed by

$$\mathrm{CSE}\,(\hat{Y}_{d,m}) = (V_c(\hat{Y}_{d,m}))^{1/2}$$

(iii) *Root Conditional Mean Squared Error*, given by

$$\mathrm{RCMSE}\,(\hat{Y}_{d,m}) = (V_c(\hat{Y}_{d,m}) + B_c^2(\hat{Y}_{d,m}))^{1/2}$$

where $\hat{Y}_{d,m}$ denotes one of the estimators studied ($m = E, R, S, C$), and the expressions of bias and variance are given respectively in Sections 5 and 6.

We observe that in this empirical study, the α value of the composite estimator has been obtained using the approximation to the optimum α in the unconditional sample space given by (Schaible 1978)

$$\alpha = MSE(\hat{Y}_{d,s})/(MSE(\hat{Y}_{d,S}) + MSE(\hat{Y}_{d,R}))$$

As seen from Table 2, in the two selected provinces, the RCB of estimator E traces an increasing curve from negative to positive values as n_d increases; furthermore, the conditional bias is very pronounced when n_d assumes the smaller and the larger values; while the RCB is near zero when n_d is close to its expected value $E(n_d)$ (that is 15,2 for Florence and 7,6 for Siena). The RCB of estimator R is essentially constant assuming a small value when $n_d \geq E(n_d)$; conversely, the RCB presents larger values when $n_d < E(n_d)$. The RCB of estimator S traces an essentially constant nonzero level over the entire n_d range: particularly for the province of Florence, the RCB values of estimator S are, generally, in the interval 0.03–0.05; while in the province of Siena the RCB values of estimator S are generally in the interval 0.15–0.18. This is due to the fact that the weight, in terms of resident population, of the strata that are composed entirely by PSUs of the province is much larger for Florence than for Siena. The RCB of estimator C is roughly constant throughout the range n_d values, with intermediate values between those of estimators S and R.

From Table 3 we can observe that CSE presents similar behaviour in the selected

Table 2. Relative conditional bias for selected provinces

n_d	Expansion	Ratio	Synthetic	Composite
		Florence		
11	−0.152	0.065	−0.036	−0.006
12	−0.122	0.039	−0.049	−0.023
13	−0.047	0.067	−0.026	0.002
14	−0.017	0.082	−0.013	0.016
15	−0.001	0.013	−0.038	−0.023
16	0.046	0.011	−0.044	−0.028
17	0.099	0.016	−0.046	−0.028
18	0.154	0.021	−0.036	−0.019
19	0.190	0.011	−0.036	−0.022
20	0.233	0.009	−0.043	−0.027
21	0.311	0.029	−0.058	−0.032
22	0.348	0.017	−0.054	−0.033
		Siena		
4	−0.889	−0.511	0.201	−0.136
5	−0.646	−0.200	0.177	−0.002
6	−0.426	−0.131	0.158	0.021
7	−0.133	−0.014	0.186	0.091
8	0.113	0.015	0.227	0.127
9	0.343	0.023	0.178	0.105
10	0.564	0.022	0.173	0.101
11	0.868	0.068	0.171	0.122
12	1.078	0.061	0.173	0.120

Table 3. Conditional standard errors for selected provinces

n_d	Expansion	Ratio	Synthetic	Composite
		Florence		
11	4,801	5,997	2,762	3,105
12	5,341	6,192	3,202	3,721
13	5,423	6,147	3,165	3,582
14	5,433	5,845	3,192	3,487
15	5,465	5,566	3,211	3,450
16	5,993	5,713	3,167	3,439
17	6,182	5,707	2,998	3,372
18	6,469	5,546	3,271	3,504
19	5,886	4,815	2,705	2,891
20	6,225	5,156	2,398	2,905
21	5,019	4,021	2,525	2,567
22	5,236	3,903	2,426	2,514
		Siena		
4	587	2,712	730	1,371
5	1,711	2,690	556	1,394
6	1,827	2,764	614	1,386
7	2,215	2,556	636	1,348
8	3,469	2,449	641	1,266
9	3,081	2,281	639	1,198
10	3,124	2,064	613	1,102
11	3,651	2,116	640	1,213
12	3,717	1,947	643	1,196

provinces. As n_d increases, the CSE of estimator E shows increasing behaviour (which underscores the less satisfactory performance of this estimator), while the CSE of estimator R decreases (as expected). The CSE of estimator S is essentially constant with lower values than those assumed by estimators R and E. The CSE of estimator C decreases slightly as n_d increases with values marginally greater than those of estimator S.

Table 4 shows that estimator S is the most efficient for all n_d values. It is followed by estimators C and R, while estimator E falls way behind the others, due in large part to a considerable conditional bias. Furthermore, we note that RCMSE of estimator E is shaped as a U curve, first decreasing and then increasing. The RCMSE of estimators R and C generally decreases, as n_d increases, while estimator S shows an essentially constant behaviour.

In conclusion, we may observe the following from the obtained results:

- Generally, estimator E has a large bias and is less efficient for most n_d values and should not be used, except when the realised sample size n_d is in the immediate vicinity of the expected value $E(n_d)$.
- Estimator R is almost conditionally unbiased, and the variance and MSE decrease as n_d increases. The estimators S and C are the most efficient estimators for all n_d values; but they present a much larger bias than that for R when n_d is higher than its expected value, while the bias of S and C is approximately equal to that of estimator R when

Table 4. Root conditional mean squared errors for selected provinces

n_d	Expansion	Ratio	Synthetic	Composite
		Florence		
11	7,594	6,505	3,092	3,113
12	7,152	6,371	3,718	3,823
13	5,721	6,674	3,321	3,583
14	5,473	6,645	3,231	3,542
15	5,465	5,589	3,535	3,562
16	6,250	5,727	3,604	3,606
17	7,296	5,740	3,488	3,537
18	8,818	5,604	3,555	3,581
19	9,450	4,834	3,038	3,013
20	10,980	5,141	3,193	3,089
21	13,064	4,175	3,382	2,852
22	14,474	3,962	3,213	2,820
		Siena		
4	5,954	4,355	1,525	1,645
5	4,636	3,002	1,302	1,394
6	3,378	2,899	1,222	1,392
7	2,386	2,557	1,393	1,479
8	3,550	2,451	1,643	1,522
9	3,815	2,286	1,351	1,387
10	4,887	2,069	1,304	1,292
11	6,842	2,164	1,309	1,462
12	8,093	1,989	1,318	1,438

$n_d < E(n_d)$. This suggests the possibility of an estimation technique based on a choice between estimators S and C when $n_d < E(n_d)$; when $n_d > E(n_d)$ estimator R is preferred.

8. Conclusions

The main contribution of this article is the derivation of the expressions of bias and variance of some relevant estimators for small areas in the conditional approach. The study is developed in the context of a two-stage sampling design stratified for PSUs with the selection of only one PSU in each stratum. This sampling design is relevant because it is used in *household surveys* conducted by the Italian National Institute of Statistics. The expressions of bias and variance given in the study allow the development of comparative analyses that aim to study the empirical properties of the estimators examined here.

The numerical results presented in this article allow the characterisation of the different conditional performances of the estimators in the functioning of the different selected sample sizes of primary units belonging to the small area. These different performances are useful in order to choose the best estimation technique for inference.

In order to compare the properties of the estimators examined here both in the conditional and in the unconditional settings, we note the following

- the ratio and composite estimators would seem to be preferable, considering just the bias: the ratio estimator is approximately unbiased in both settings; the composite estimator has a low bias. The expansion estimator, that is unbiased in the unconditional setting, has a large conditional bias. The synthetic estimators are characterised by large bias in both settings;
- considering both bias and the MSE, the composite estimator would seem to be preferable, since as is well-known, it is characterised by low values of MSE and of the bias in the unconditional setting, and as shown by the experimental results obtained here, it has good performances in the conditional setting; furthermore, as shown when $n_d > E(n_d)$ estimator R is generally the best estimator in the conditional setting, that is the estimator characterised by the lowest values of the conditional MSE.

Finally, we observe that the conditional properties of the examined estimators are similar to those given in the articles by Särndal and Hidiroglou (1989) and by Russo and Falorsi (1993) developed in the context of simple random sampling and simple two stage sampling respectively.

9. References

Cochran, W.G. (1977). Sampling Techniques. New York: John Wiley.

Gosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 9, 55–93.

Holt, D. and Smith, T.M.F. (1979). Post-Stratification. Journal of the Royal Statistical Society, Part A, 142, 33–46.

Rao, J.N.K. (1985). Conditional Inference in Survey Sampling. Survey Methodology, 11, 15–31.

Russo, A. and Falorsi, P.D. (1993). Conditional and Unconditional Properties of Small Area Estimators in Two Stage Sampling. Invited papers of the International Scientific Conference: Small Area Statistics and Survey Design, Warsaw, 30 September–3 October, 1992, 251–270.

Royall, R.M. and Cumberland, W.G. (1985). Conditional Coverage Properties of Finite Population Confidence Intervals. Journal of the American Statistical Association, 80, 355–359.

Särndal, C.E. and Hidiroglou, M.A. (1989). Small Domain Estimation: A Conditional Analysis. Journal of the American Statistical Association, 84, 266–275.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New-York: Springer-Verlag.

Schaible, W.L. (1978). Choosing Weights for Composite Estimators for Small Area Statistics. Proceedings of the American Statistical Association, Survey Research Section, 741–746.

Singh, M.P., Gambino, J., and Mantel, H.J. (1994). Issues and Strategies for Small Area Data. Survey Methodology, 20, 3–22.

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

Received July 1997
Revised August 1998