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 approach is based on 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 Rome, Italy. E-mail: falorsi@istat.it
2 Third University of Rome, Via Segre 2, Roma, Italy. E-mail: alдорuss@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.

© Statistics Sweden
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 \( h \)th PSU ;
- \( M_h \) = set of SSUs belonging to the \( N_h \) PSUs of the stratum \( h \);
- \( Y_{hi} \) value of the variable of interest \( y \) in the SSU \( hj \).

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, ..., \( n \).
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)

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

(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 $h_i$ is defined by (Särndal, Swensson, and Wretman 1992, p. 31)

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

(3)

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

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 $n_{d}$ 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:

<table>
<thead>
<tr>
<th>Stratum</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>h</th>
<th>...</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic configuration</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

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 $\delta_h$ ($h = 1, \ldots, n$) 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

$$p(g_w) = \prod_{h=1}^{n} \pi_{g_{h}}^{b_{h}} (1 - Z_{h})^{1-b_{h}}$$

(4)

where

$$Z_{h} = \sum_{j=1}^{n} \pi_{h_j}$$

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

Consequently, we have

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

(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 \( h \’autres \); \( C_h \) 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

\[
p(g_v(hi)) = \pi_{hi} \prod_{u=1}^{n-1} Z_u^h (1 - Zu)^{1-\delta_u} \quad (u \neq h)
\]  

Consequently, we have

\[
\sum_{s(hi) \in U_c} p(s(hi)) = \sum_{v=1}^{n-1} p(g_v(hi))
\]  

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

\[
\sum_{s(hi) \in U_c} p(s(hi)) = \sum_{v=1}^{n-1} p(g_v(hi))
\]  

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

\[
N_{d1} N_{d2} N_{d3} (N_4 - N_{d4}) (N_5 - N_{d5})
\]  

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

\[
p(g_1) = Z_1 Z_2 Z_3 (1 - Z_4) (1 - Z_5)
\]

\[
= \left( \sum_{i=1}^{N_{d1}} \pi_1^i \right) \left( \sum_{i=1}^{N_{d2}} \pi_2^i \right) \left( \sum_{i=1}^{N_{d3}} \pi_3^i \right) \left( 1 - \sum_{i=1}^{N_{d4}} \pi_4^i \right) \left( 1 - \sum_{i=1}^{N_{d5}} \pi_5^i \right)
\]  

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

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

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

$$
\sum_{s \in U_c} p(s) = \sum_{h_1=1}^{n-n_d} \sum_{h_2 > h_1}^{n-(n_d-1)} \sum_{h_3 > h_2}^{n-(n_d-2)} Z_{h_1} Z_{h_2} \cdots Z_{h_{n_d}} (1 - Z_{q_1}) \cdots (1 - Z_{q_{n-n_d}}) \tag{11}
$$

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 $h_i$ 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 $1$ belonging to the small area is selected, is given by

$$
p(g_1(1)) = \pi_{11} Z_{11} Z_{12} (1 - Z_4) (1 - Z_3)
$$

For $p(g_2(1)), p(g_3(1)), \ldots, p(g_6(1))$, these expressions are similar to (13), in which the...
first factor equals $\pi_{1h}$. Hence, using expression (7) we have
\[
\sum_{s(1i) \in U_i} p(s(1i)) = \sum_{s(1i) \in U_i} \frac{6}{\pi_{1h}} \sum_{h_1 = 2}^{4} \sum_{h_2 > h_1}^{5} Z_{h1} Z_{h2} (1 - Z_{q1}) (1 - Z_{q2})
\]
(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 $\tilde{h}$ the stratum under study ($\tilde{h} = 1$, or $\tilde{h} = 2, \ldots, \tilde{h} = 5$), we then rearrange the stratum codes giving to generic stratum $h(h \neq \tilde{h})$ the code $\gamma$ expressed by
\[
\gamma = \begin{cases} h & \text{for } h < \tilde{h} \\ h - 1 & \text{for } h > \tilde{h} \end{cases}
\]
(15)
Thus, the probability expressed by (7) is given by
\[
\sum_{s(\tilde{h}i) \in U_i} p(s(\tilde{h}i)) = \pi_{\tilde{hi}} \sum_{\gamma_1 = 1}^{n-(n_1-2)} \sum_{\gamma_2 > \gamma_1}^{n} Z_{\gamma_1} Z_{\gamma_2} (1 - Z_{q1}) (1 - Z_{q2})
\]
\[= \pi_{\tilde{hi}} \sum_{\gamma_1 = 1}^{n-(n_1-2)} \sum_{\gamma_2 > \gamma_1}^{n} Z_{\gamma_1} Z_{\gamma_2} (1 - Z_{q1}) (1 - Z_{q2})
\]
(16)
where $(\gamma_1, \gamma_2)$ are stratum indexes, expressed by (15), in which the sample PSU belongs to small area $d$, and $q_1, q_2$ 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
\[
\sum_{s(\tilde{h}i) \in U_i} p(s(\tilde{h}i)) =
\]
\[= \pi_{\tilde{hi}} \sum_{\gamma_1 = 1}^{n-(n_d-2)} \sum_{\gamma_2 > \gamma_1}^{n-(n_d-3)} \sum_{\gamma_{n_d-i-1} > \gamma_{n_d-i-2}}^{n} Z_{\gamma_1} Z_{\gamma_2} \cdots Z_{\gamma_{n_d-i-1}} (1 - Z_{q1}) \cdots (1 - Z_{q_{n-n_d}})
\]
(17)
with the PSU $\tilde{h}i$ belonging to small area $d$, $(\gamma_1, \gamma_2, \ldots, \gamma_{n_d-i-1}) \neq (q_1, \ldots, q_{n-n_d})$ 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 $\tilde{h}i$ does not belong to the small area $d$; we have
\[
\sum_{s(\tilde{h}i) \in U_i} p(s(\tilde{h}i))
\]
\[= \pi_{\tilde{hi}} \sum_{\gamma_1 = 1}^{n-(n_d-1)} \sum_{\gamma_2 > \gamma_1}^{n-(n_d-2)} \sum_{\gamma_{n_d-i-1} > \gamma_{n_d-i-2}}^{n} Z_{\gamma_1} Z_{\gamma_2} \cdots Z_{\gamma_{n_d-i}} (1 - Z_{q1}) \cdots (1 - Z_{q_{n-n_d-i-1}})
\]
(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
\[ \hat{Y}_{d,E} = \sum_{h \in D} \sum_{j \in m_h} Y_{hij} \hat{\delta}_{hi}(\pi_{j,hi} \pi_{hi})^{-1} \]  
(19)
\[ \hat{Y}_{d,R} = (\hat{Y}_{d,E} / \hat{X}_{d,E}) X_d \]  
(20)
\[ \hat{Y}_{d,S} = (\hat{Y}_{d,E} / \hat{X}_E) X_d \]  
(21)
\[ \hat{Y}_{d,C} = \alpha \hat{Y}_{d,R} + (1 - \alpha) \hat{Y}_{d,S} \]  
(22)
being
\[ \hat{X}_{d,E} = \sum_{h \in D} \sum_{j \in m_h} X_{hij} \hat{\delta}_{hi}(\pi_{j,hi} \pi_{hi})^{-1}, \quad \hat{Y}_E = \sum_{h \in D} \sum_{j \in m_h} Y_{hij} \hat{\delta}_{hi}(\pi_{j,hi} \pi_{hi})^{-1} \]
\[ \hat{X}_E = \sum_{h \in D} \sum_{j \in m_h} X_{hij} \pi_{j,hi} \pi_{hi}^{-1}, \quad X_d = \sum_{h \in D} \sum_{i \in N_h} \sum_{j \in m_{hi}} X_{hij} \]
where: \( \pi_{j,hi} = m_{hi}/M_{hi} \) is the inclusion probability of the SSU \( hij \) conditional on the selection of the PSU \( hi \); \( \hat{\delta}_{hi} \) 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 \( d \); \( \alpha \) is a constant \( 0 \leq \alpha \leq 1 \). Overviews of options in the choice of \( \alpha \) 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 \( \alpha \). 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 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
\[ \hat{Y}_{d,E} = \sum_{h \in D} \sum_{i \in N_{hi}} (1/\pi_{hi}) \lambda_{hi} \sum_{j \in M_{hi}} (1/\pi_{j,hi}) Y_{hij} \lambda_{hi} \]  
(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
\[ E_c(\hat{Y}_{d,E}) = \sum_{h \in D} \sum_{i \in N_{hi}} (1/\pi_{hi}) E_{1c}(\lambda_{hi}) \sum_{j \in M_{hi}} (M_{hi}/m_{hi}) Y_{hij} E_2(\lambda_{hi}) \]
= \[ \sum_{h \in D} \sum_{i \in N_{hi}} (\pi_{c,hi}/\pi_{hi}) Y_{hi} \]  
(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

\[ B_c(\tilde{Y}_{d,E}) = E_c(\tilde{Y}_{d,E}) - Y_d = \sum_{h \in D} \sum_{i \in N_{dh}} (\pi_{c,hi}/\pi_{hi}) Y_{hi} - Y_d \]  

(25)

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

\[ \tilde{Y}_{d,R} = \left[ E_c(\tilde{Y}_{d,E})/E_c(\tilde{X}_{d,E}) \right] X_d + \left[ X_d/E_c(\tilde{X}_{d,E}) \right] \left[ \tilde{Y}_{d,E} - E_c(\tilde{Y}_{d,E}) \right] + X_d \left[ E_c(\tilde{Y}_{d,E})/E_c(\tilde{X}_{d,E}) \right] \left[ \tilde{X}_{d,E} - E_c(\tilde{X}_{d,E}) \right] \]  

(26)

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

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

(27)

Therefore, the conditional bias of the ratio estimator is

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

(28)

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

\[ B_c(\tilde{Y}_{d,S}) = \left[ E_c(\tilde{Y}_{E})/E_c(\tilde{X}_{E}) \right] X_d - Y_d \]  

(29)

where

\[ E_c(\tilde{X}_{E}) = \sum_{h \in D} \sum_{i \in N_h} (\pi_{c,hi}/\pi_{hi}) X_{hi} \]

\[ E_c(\tilde{Y}_{E}) = \sum_{h \in D} \sum_{i \in N_h} (\pi_{c,hi}/\pi_{hi}) Y_{hi} \]

Hence the conditional bias of the composite estimator is given by

\[ B_c(\tilde{Y}_{d,C}) = \alpha B_c(\tilde{Y}_{d,R}) + (1 - \alpha) B_c(\tilde{Y}_{d,S}) \]  

(30)

6. Variance

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

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

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

\[ V_c(\tilde{Y}_{d,E}) = \sum_{h \in D} \left[ \sum_{i \in N_{dh}} (Y_{hi}/\pi_{hi})^2 \pi_{c,hi}(1 - \pi_{c,hi}) - 2 \sum_{i > i} Y_{hi} Y_{hi'} \pi_{c,hi} \pi_{c,hi'} (\pi_{hi} \pi_{hi'})^{-1} \right] \]
\[ + \sum_{i \in N_d} (\pi_{c,h}/\pi_{hi}) [M_{hi} - m_{hi}] [\pi_{j,hi}(M_{hi} - 1)]^{-1} \sum_{j \in M_{hi}} (Y_{hi} - (Y_{hi}/M_{hi}))^2 \\
+ 2 \sum_{h \neq h} \sum_{i \in N_d} \sum_{i' \in N_{d,i'}} \left( \pi_{c,h,i} - \pi_{c,h,i'} \pi_{c,h,i'} \right) Y_{hi} Y_{i'hi} (\pi_{hi} \pi_{hi'})^{-1} \left] \right) \]  

in which

\[ \pi_{c,h,i'} = \sum_{s(h,i',s) \in U_{hi}} p_{c}(s(h,i',s)) = \left( \sum_{s \in U_{hi}} p(s) \right)^{-1} \sum_{s(h,i',s) \in U_{hi}} (p(s(h,i',s))) \]

denotes the conditional second order inclusion probabilities of the primary units \( h_i \) and \( h_{i'} \), being \( s(h,i',s) \) the generic sample of \( U_{hi} \) that contains these PSUs and \( p(s(h,i',s)) \) the unconditional selection probability of the sample \( s(h,i',s) \).

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

\[ Z_{hi} = [X_d/E_c(X_{d,E})] [Y_{hi} - (E_c(\hat{Y}_{d,E})/E_c(X_{d,E}))X_{hi}] \]

\[ Z_{i'} = \sum_{j \in M_{hi}} Z_{i'j} \]

The conditional variance of the synthetic estimator is given by

\[ V_c(\hat{Y}_{d,E}) = \sum_{h \in D} \left[ \sum_{i \in N_h} (Q_{hi}/\pi_{hi})^2 \pi_{c,h,i}(1 - \pi_{c,h,i}) - 2 \sum_{i > i'} Q_{hi} \pi_{c,h,i} \pi_{c,h,i'} (\pi_{hi} \pi_{hi'})^{-1} \right. \]

\[ + \sum_{i \in N_h} \left( \pi_{c,h,i}/\pi_{hi} \right) [M_{hi} - m_{hi}] [\pi_{j,hi}(M_{hi} - 1)]^{-1} \sum_{j \in M_{hi}} (Q_{hj} - (Q_{hj}/M_{hi}))^2 \]

\[ + 2 \sum_{h \neq h} \sum_{i \in N_h} \sum_{i' \in N_{h,i'}} \left( \pi_{c,h,i} - \pi_{c,h,i'} \pi_{c,h,i'} \right) Q_{hi} Q_{i'hi} (\pi_{hi} \pi_{hi'})^{-1} \left] \right) \]

where

\[ Q_{hi} = [X_d/E_c(X_{d,E})] [Y_{hi} - (E_c(\hat{Y}_{d,E})/E_c(X_{d,E}))X_{hi}] \]

\[ Q_{hi} = \sum_{j \in M_{hi}} Q_{hi} \]

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

\[ W_{hi} = \alpha \delta_{hi} Z_{hi} + (1 - \alpha) Q_{hi} \]

\[ W_{hi} = \sum_{j \in M_{hi}} W_{hi} \]

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 illus-
tration 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 pos-
sible configurations as described in Section 3; consequently, we have obtained the condi-
tional 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 prob-
babilities $\pi_{c,hi}$ we have calculated, for each $n_d$, the following conditional performance
measures:

(i) **Relative Conditional Bias**, defined as

$$
\text{RCB} \left( \hat{Y}_{d,m} \right) = \frac{B_c(\hat{Y}_{d,m})}{Y_d}
$$

(ii) **Conditional Standard Error**, expressed by

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

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

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

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 \( \alpha \) value of the composite estimator has been obtained using the approximation to the optimum \( \alpha \) in the unconditional sample space given by (Schaible 1978)

\[
\alpha = \frac{\text{MSE}(\hat{Y}_{d,s})}{\text{MSE}(\hat{Y}_{d,s}) + \text{MSE}(\hat{Y}_{d,R})}
\]

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

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

\(n_d < E(n_d)\). 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 $n_j > E(n_j)$ 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


Received July 1997
Revised August 1998