Survey Methodology, June 2003
Vol. 29, No. 1, pp. 53-61
Statistics Canada, Catalogue No. 12-001-XPB

Conditional and Unconditional Analysis of Some Small Area Estimators in Complex Sampling

Loredana Di Consiglio, Piero Demetrio Falorsi, Stefano Falorsi and Aldo Russo [1]

Abstract

This work deals with the unconditional and conditional properties of some well known small area estimators: expansion, post-stratified ratio, synthetic, composite, sample size dependent and the empirical best linear unbiased predictor. As it is commonly used in household surveys conducted by the National Statistics Institute of Italy, a two-stage sampling design is considered. An evaluation is carried out through a simulation based on 1991 Italian Census data. The small areas considered are the Local Labour Market Areas, which are unplanned domains that cut across the boundaries of the design strata.

Key Words: Relative conditional bias; Relative root conditional MSE; Conditional coverage rate.

1. Introduction

Sampling theorists prefer to plan the sampling strategy on the basis of the unconditional sample space U_u, *i.e.*, the set of all possible samples (*unconditional approach*). However, after data collection, the reliability of an estimate obtained by means of an estimator $\tilde{Y}$, can be evaluated either unconditionally or conditionally; *i.e.*, the evaluation can be assessed on the conditional sample space U_C (*conditional approach*), where U_C is the set of samples with 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 papers consider the case of simple random sampling. In the context of small area estimation, the conditional and unconditional properties of some estimators for a two-stage sampling design with stratification of the primary sampling units have been studied in Russo and Falorsi (1993), Russo and Falorsi (1996), Falorsi and Russo (1999) and Falorsi, Falorsi and Russo (2000).

This paper considers a two-stage sampling design with stratification of the Primary Sampling Units (PSUs). This kind of design is generally used in household surveys conducted by the National Statistics Institute, *e.g.*, the Labour Force Survey (LFS). The aim of this work is to evaluate, on the basis of a simulation study, the conditional and unconditional properties of some important small area estimators.

The principal aspects of our investigation are:

- the simulation study is based on a sample design with strata, cluster delineation and sample size similar to those used in the LFS;
- the small areas considered are the Local Labour Market Areas (LLMAs), which are unplanned domains that cut across the boundaries of the design strata;
- the conditional analysis is developed using a sample space U_C, as reference set, consisting of all the possible samples containing a fixed number of PSUs belonging to the LLMA;
- the estimators examined are expansion, post-stratified ratio, synthetic, composite, sample size dependent and empirical best linear unbiased predictor. For a review see Ghosh and Rao (1994), Singh, Gambino and Mantel (1994), Pfeffermann (1999) and Rao (1999).

In section 2 the sampling design, the parameters of interest and the current estimator used by the LFS are described. Section 3 illustrates the small area estimators examined in the present work. In section 4 the empirical results of the simulation study are shown. Section 5 contains a short summary with suggestions for extension of the analysis.

2. Description of the LFS Sampling Strategy

2.1 Sample Design

The LFS is a quarterly sample of about 72,000 households designed to produce estimates of the labour force status of the population at national and regional levels. The survey in each quarter is based on a composite design. Within a given province (administrative area inside the region) the municipalities are divided into two area types:

1. Loredana Di Consiglio, Piero Demetrio Falorsi and Stefano Falorsi, Istituto Nazionale di Statistica, Via Cesare Balbo, 16 - 00184 Roma, Italy; Aldo Russo Università di Roma TRE Via C. Segre, 2-00142 Roma, Italy.

the Self-Representing Area (SRA) – consisting of the larger municipalities – and the Non Self-Representing Area (NSRA) – consisting of the smaller ones.

In the SRA a stratified cluster sampling design is applied. Each municipality is a single stratum and the PSUs are the households selected by means of systematic sampling. All members of each sampled household are interviewed.

In the NSRA the sample is based on a stratified two-stage sample design. The PSUs are the municipalities, while the Secondary Sampling Units (SSUs) are the households. The PSUs are divided into strata of the same magnitude in terms of population size. Two sample PSUs are selected from each stratum without replacement and with probability proportional to the PSU's population size. The SSUs are selected by means of systematic sampling in each PSU. All members of each sample household are interviewed.

2.2 Notation and Parameter of Interest

For simplicity's sake we will introduce notation only for the two-stage sampling design of the NSRA. Note that the derivation of the quantities and expressions for the SRA case is a special case of NSRA.

With reference to the generic geographical region we introduce the following subscripts: p ($p = 1, \ldots, L$) for province; h ($h = 1, \ldots, H_p$) for stratum; i for municipality; j for household; a ($a = 1, \ldots, A$) for age-sex group. A quantity associated to stratum h, municipality i, and household j will be briefly referred to as a quantity in hij; a quantity associated to stratum h and municipality i will be referred to as a quantity in hi. The following notation is also used: N_h for the number of municipalities in h; P_h for the number of persons in h; n_h for the number of sample municipalities in h; M_{hi} for the number of households in hi; P_{hi} for the number of persons in hi; m_{hi} for the number of sample households in hi; P_{ahij} for the number of persons in group a belonging to hij and P_{hij} for the number of persons in hij.

Further let

$$Y = \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ahij}$$

be the total of the characteristic y for the regional population, where Y_{ahij} denotes the total of the characteristic of interest y for the P_{ahij} persons in group a in household hij.

2.3 Estimator of Y

An estimate of total Y is obtained by means of a post-stratified ratio estimator expressed by

$$\hat{Y}^R = \sum_{a=1}^{A} \frac{\hat{Y}_a^E}{\hat{P}_a^E} P_a \qquad (1)$$

where

$$\hat{Y}_a^E = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij}\, Y_{ahij}$$

and

$$\hat{P}_a^E = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij}\, P_{ahij} \qquad (2)$$

represent unbiased estimators of

$$Y_a = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ahij}$$

and

$$P_a = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} P_{ahij}.$$

The symbol K_{hij}, that denotes the *basic weight*, is expressed by (Cochran 1977)

$$K_{hij} = \frac{P_h}{n_h P_{hi}} \frac{M_{hi}}{m_{hi}}.$$

Note that for the SRA

$$n_h = 1 \quad \text{and} \quad P_{hi} = P_h, \quad \text{so} \quad K_{hij} = \frac{M_{hi}}{m_{hi}}.$$

3. Small Area Estimators

We now consider the problem of estimating the total of a y variable for units belonging to a small area. Let d ($d = 1, \ldots, D$) be the generic small area of a given geographical region. Since the LLMAs may cut across provinces, the total of interest in small area d is defined by

$$Y_d = \sum_{a=1}^{A} Y_{da} \qquad (3)$$

with

$$Y_{da} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{N_{dh}} \sum_{j=1}^{M_{hi}} Y_{ahij}$$

where L_d denotes the provinces including part of the small area d, H_{dp} are the strata of province p intersecting the small area d and N_{dh} denotes the municipalities of stratum h belonging to small area d.

The choice of an estimation method basically depends on available information. In Italy the accessible information at the small area level is currently very poor: only total persons in age-sex groups can be obtained at the municipality level; this is why all the small area estimators considered here will be based on this information only. In the simulation work we have considered the following *direct estimators*:

(i) the *expansion estimator*

$$\hat{Y}_d^E = \sum_{a=1}^{A} \hat{Y}_{da}^E \qquad (4)$$

where

$$\hat{Y}_{da}^E = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_{dh}} \sum_{j=1}^{m_{hi}} K_{hij}\, Y_{ahij}$$

is the expansion estimator of Y_{da} and n_{dh} is the number of sampled municipalities of stratum h belonging to LLMA d;

(ii) the *post-stratified ratio estimator*

$$\hat{Y}_d^R = \sum_{a=1}^{A} \frac{\hat{Y}_{da}^E}{\hat{P}_{da}^E} P_{da} \qquad (5)$$

in which

$$\hat{Y}_{da}^E = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_{dh}} \sum_{j=1}^{m_{hi}} K_{hij}\, Y_{ahij},\ \hat{P}_{da}^E = \sum_{l=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_{dh}} \sum_{j=1}^{m_{hi}} K_{hij},$$

$$P_{da} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{pd}} \sum_{i=1}^{N_{dh}} \sum_{j=1}^{M_{hi}} P_{ahij}.$$

In the simulation work reported here we have considered the following *design-based indirect estimators*:

(iii) the *synthetic estimator*

$$\hat{Y}_d^S = \sum_{a=1}^{A} \frac{\hat{Y}_a^E}{\hat{P}_a^E} P_{da} \qquad (6)$$

in which $\hat{Y}_a^E$ and $\hat{P}_a^E$ are expressed by formulas (2). The estimator (6) is based on the underlying assumption that, for each post-stratum a, the small area mean equals the mean at the regional level;

(iv) the *composite estimator*, considered in two alternative forms

$$\hat{Y}_d^{C1} = \alpha_d\, \hat{Y}_d^R + (1-\alpha_d)\, \hat{Y}_d^S \qquad (7)$$

$$\hat{Y}_d^{C2} = \alpha\, \hat{Y}_d^R + (1-\alpha)\, \hat{Y}_d^S \qquad (8)$$

where $\alpha_d\,(0 \le \alpha_d \le 1)$ is a specific small area weight while $\alpha\,(0 \le \alpha \le 1)$ is a common weight for all the LLMAs of the region. The methods used to calculate weights α_d and α will be described in subsection 4.1. Both of the composite estimators equal by definition the synthetic estimator when the sample size in the small area equals zero;

(v) the *sample size dependent estimator* (SSD), expressed by

$$\hat{Y}_d^{\mathrm{SD}} = w_d\, \hat{Y}_d^R + (1-w_d)\, \hat{Y}_d^S \qquad (9)$$

where

$$w_d = \begin{cases} 1 & \text{if } \hat{P}_d^E \ge \lambda\, P_d \\ \hat{P}_d^E \,/\, (\lambda\, P_d) & \text{otherwise} \end{cases}$$

where λ is a given constant, $\hat{P}_d^E = \sum_{a=1}^{A} \hat{P}_{da}^E$ and $P_d = \sum_{a=1}^{A} P_{da}$.

The estimator (9) is based on the result that the performance of the post-stratified ratio estimator depends on the proportion of the sample falling in the small area. If the proportion of the sample within the small area is reasonably large then the estimator (9) equals the post-stratified ratio estimator. Otherwise it becomes a composite estimator with increasing weight $(1-w_d)$ on the synthetic estimator, as the size of the sample in the small area decreases.

Finally, in the framework of *model-based indirect predictors*, we consider:

(vi) the *empirical best linear unbiased predictor (EBLUP)*

$$\hat{Y}_d^{EP} = \gamma_d \hat{Y}_d^R + (1-\gamma_d)\, x_d' \tilde{\beta} \qquad (10)$$

where

$$\tilde{\beta} = \left[\sum_{d=1}^{D} x_d x_d' \Big/ (\tilde{\sigma}_v^2 + \psi_d) \right]^{-1} \left[\sum_{d=1}^{D} x_d \hat{Y}_d^R \Big/ (\tilde{\sigma}_v^2 + \psi_d) \right],$$

$$\gamma_d = \tilde{\sigma}_v^2 \,/\, (\tilde{\sigma}_v^2 + \psi_d) \qquad (11)$$

that is based on the well-known area level linear mixed model of Fay and Herriot (1979):

$$\hat{Y}_d^R = x_d' \beta + v_d + e_d \qquad (12)$$

in which: β is the vector of regression parameters, x_d is a vector of area-specific auxiliary data, v_d are uncorrelated random area effects with mean zero and variance σ_v^2, e_d are independent sampling errors with mean zero and known variance ψ_d, $\tilde{\beta}$ is the weighted least squares estimator of β with weights $(\sigma_v^2 + \psi_d)^{-1}$ and $\tilde{\sigma}_v^2$ is suitable estimator of σ_v^2. In this work we utilise an asymptotically consistent estimator of σ_v^2 that can be obtained iteratively by alternating weighted least squares estimation for β with the solution of

$$\frac{\sum_{d=1}^{D} (\hat{Y}_d^R - x_d'\, \beta)^2}{\sigma_v^2 + \psi_d} = D - k$$

for σ_v^2, where k is the number of elements of vector x_d, corresponding to the number of auxiliary variables in the model (12). The previous description is based on the assumption that the variances ψ_d are known; in practice these variances are seldom known. In the present study we have considered two different

methods (see subsection 4.1) for evaluating sampling variances. From these two methods we obtain two alternative empirical best linear unbiased predictors, $\hat{Y}_d^{EP1}$ and $\hat{Y}_d^{EP2}$.

4. Empirical Study

4.1 Simulation of the LFS Sample Design

In order to illustrate the conditional and unconditional properties of the estimators discussed in the preceding section, we carried out a simulation study involving repeated draws of a sample design with strata and cluster delineation and sample size similar to those used in LFS. The study can be summarised as follows:

- the information referring to the auxiliary variables and the totals of interest $Y_d\,(d=1,\ldots,D)$ are taken from the 1991 General Population Census of Italy;
- the variables of interest are Employed, Unemployed and persons searching for their first job;
- the auxiliary variables for the post-stratification of the members of the sampling households are sex and age;
- the small areas of interest are the 27 LLMAs of the Lazio region;
- for the Monte Carlo simulation $R = 2{,}000$ two-stage LFS samples were selected for each one of the five provinces of the Lazio region;
- the number of sex-age classes considered in the construction of the synthetic estimators equals 28; the age groups are 0–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, more than 74;
- the SSD estimator has been evaluated with different values for the parameter $\lambda\,(\lambda = 2/3, \lambda = 1.5$ and $\lambda = 2)$; the best performance in terms of mean square error has been obtained for $\lambda = 2$, so in this work only the results for SSD with $\lambda = 2$ are reported;
- for the empirical best linear unbiased predictors $\hat{Y}_d^{EP1}$ and $\hat{Y}_d^{EP2}$ we have removed from the analysis the LLMA of Rome. In fact the LLMA of Rome is very big in terms of population and we have verified that it has too much influence in the model. The model has been fitted separately for two groups of small areas (see section 5.1 for the definition of groups). The following covariates have been chosen:
  1) in the model for Employed and Unemployed, the province (administrative area contained in region) and the number of persons in age groups 14–35 and 35–65 by sex;
  2) in the model for persons searching for their first job, the province and the number of persons in age groups 14–25 and 25–35 by sex.

  The reduction of the number of classes with respect to the synthetic case was necessary because the number of small areas in this study is not large enough:
- the weights of composite estimator $\hat{Y}_d^{C1}$ correspond to the optimal weights given by the ratio of the MSE of the synthetic estimator over the sum of the variance of the direct estimator and the MSE of synthetic estimator (Schaible 1978). These quantities were actually evaluated on the 1991 census data;
- the unique regional weight of composite estimator $\hat{Y}_d^{C2}$ is the estimated optimal one for the average MSE of the composite estimators of all areas (Purcell and Kish 1979) given by

$$\alpha = 1 - \frac{\sum_{d=1}^{D} \hat{\mathrm{var}}(\hat{Y}_d^R)}{\sum_{d=1}^{D} (\hat{Y}_d^S - \hat{Y}_d^R)^2}.$$

  The resulting estimator is sample dependent. We have not pursued this method for small area specific weights due to the high variability of each area MSE and variance estimation. A smoothed model has been used to improve the stability of the evaluation of variances: the variance for the SRAs is obtained applying standard formulas for variance estimation on the linearized variables. For the NSRAs the variance is obtained applying a common design effect evaluated at the regional level to the simple random sampling variance estimate;
- in the predictor $\hat{Y}_d^{EP1}$, the sampling error variance ψ_d has been evaluated using census data; for predictor $\hat{Y}_d^{EP2}$ we have considered the alternative case in which ψ_d has to be evaluated through sample data: a regression model based on twelve simulated LFS samples was fitted and then the value of ψ_d predicted through the model.

4.2 Performance Measures

4.2.1 Overall Unconditional Measures

The following unconditional performance measures were calculated to assess the bias and the MSE of the estimators over the 2,000 replications and over all the *D* small areas:

- Percentage Average Absolute Relative Bias (AARB);
- Percentage Average Relative Root Mean Square Error (ARRMSE), expressed respectively by formulas

$$\mathrm{AARB}(\hat{Y}^T) = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{R}\left[\frac{\hat{Y}_d^T(r) - Y_d}{Y_d}\right]\right| 100$$

$$\text{ARRMSE}(\hat{Y}^T) = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{R}\left(\sum_{r=1}^{R}\left[\frac{\hat{Y}_d^T(r)-Y_d}{Y_d}\right]^2\right)}100$$

in which $\hat{Y}_d^T(r)$ indicates the value of the generic small area estimator T (described in section 3) obtained in the r^{th} of the R = 2,000 samples.

The same measures were also considered averaging only on subsets of small areas, with D replaced by the cardinality of the subset. For the definition of the subsets see section 5.1.

4.2.2 Conditional Measures

For each small area d, the 2,000 repeated samples were distributed over the different values of the realised number n_d of sampled municipalities belonging to small area d. For each value of n_d and for each small area d, the conditional performance measures were computed over that subset of the 2,000 samples for which the small area sample PSU count was exactly n_d.

The following conditional performance measures were considered:

– Percentage Relative Conditional Bias (RCB);

– Percentage Relative Root Conditional MSE (RCMSE);

– Conditional Coverage Rate (CCR).

These measures were calculated in the following way:

$$\text{RCB}(\hat{Y}_d^T) = \frac{1}{R_d}\sum_{r=1}^{R_d}\left[\frac{\hat{Y}_d^T(r)-Y_d}{Y_d}\right]100$$

$$\text{RRCMSE}(\hat{Y}_d^T) = \sqrt{\frac{1}{R_d}\sum_{r=1}^{R_d}\left[\frac{\hat{Y}_d^T(r)-Y_d}{Y_d}\right]^2}\ 100$$

$$\text{CCR}(\hat{Y}_d^T) = \left(\frac{1}{R_d}\sum_{r=1}^{R_d} I(r)\right)\ 100$$

in which R_d indicates the number of samples for which the PSU sample count in the small area d equals the fixed number n_d; $I(r)=1$ if the r^{th} confidence interval based on $\hat{Y}_d^T(r)$ contains the true value Y_d and $I(r)=0$ otherwise. The nominal value equals 95% and the confidence interval is the normal confidence interval where we have used as evaluation of variance the value resulting from the 2,000 replications.

5. Analysis of the Results

5.1 Unconditional Analysis

The LLMAs analysed in the simulation with their characteristics in terms of population, number of municipalities and number of LFS strata intersected are reported in Table 1. The small areas have been grouped on the basis of the ranking of the proportion of LLMA's population over the total regional population. The percent proportion of the first group ranges from 0.12% to 1.73%; the group is composed of 19 LLMAs. The percent proportion of the second group ranges from 1.9% to 5.05%; the group is composed of 7 LLMAs. The third group consists of the largest LLMA representing a percent proportion equal to 64%. The LLMAs are divided into these three groups because we expect the MSE to be larger for those LLMAs with smaller sample size.

Table 1
Local Labour Market Area (LLMA), Population, Percent Population, Number of Municipalities and Number of LFS Strata Intersected by the LLMA

LLMA	Population	Population %	Nunber Municipalities	Number Strata
398	6,005	0.12	5	2
396	7,364	0.14	3	2
391	8,901	0.17	4	2
407	11,392	0.22	2	2
393	12,500	0.24	3	3
414	12,656	0.25	4	2
406	13,051	0.25	3	2
395	16,012	0.31	5	3
390	19,823	0.39	8	3
411	23,226	0.45	5	2
394	30,193	0.59	6	3
408	45,274	0.88	5	2
392	51,789	1.01	13	5
416	59,512	1.16	10	4
402	71,906	1.40	15	5
401	72,080	1.40	34	8
400	72,235	1.41	4	3
412	78,249	1.52	5	3
409	88,984	1.73	7	4
399	97,680	1.90	42	5
405	114,361	2.23	3	2
397	133,303	2.60	18	5
413	146,133	2.85	41	5
410	170,945	3.33	6	4
404	198,010	3.86	16	8
415	259,382	5.05	35	7
403	3,314,237	64.54	65	13

In Table 2 we present the values of the unconditional performance measures AARB and ARRMSE for one of the three LFS characteristics studied: the number of Unemployed. This variable has been chosen since it is one of the most important characteristic produced by the LFS.

Table 2
Percentage Average Absolute Relative Bias and Percentage Average Root Relative Mean Square Error of the Estimators of Unemployed

Estimator	AARB	ARRMSE
Expansion	2.67	96.07
Post stratified ratio	26.20	58.29
Synthetic	18.10	19.40
Composite C1	15.52	17.34
Composite C2	8.94	31.48
SSD	10.14	29.84
EBLUP EB1*	13.36	66.57
EBLUP EB2*	12.98	74.88

* The averages for the EBLUPs do not include LLMA = 403.

Table 3 reports the same measures for each of the three previously defined groups of LLMAs.

From the analysis of the results in Tables 2 and 3, the following conclusions emerge:

- with the exclusion of the direct estimator, the bias of composite estimator $\hat{Y}^{C2}$ is almost always the smallest, or among the smaller ones, and it is very close to the bias of the SSD estimator;
- composite estimator $\hat{Y}^{C1}$ is almost always the best in terms of ARRMSE; its performance is similar to that of the synthetic estimator when taking account of the overall measure. This is due to the fact that the optimal weights are close to zero on many of the small areas considered in the simulation (note that many small areas have a percentage population under 2%). This can be confirmed by examining the results for Group 1 where the similarity of the two estimators is evident;
- the overall bias of the post-stratified ratio estimator is very high; this can be explained by the very high bias of the estimator for the areas belonging to Group 1, where the typical sample size is small;
- the model used for the empirical best linear unbiased predictors does not seem adequate, likely because we are far from the hypothesis of unbiasedness for the direct component (post-stratified ratio estimator) and due to the choice of the auxiliary variables; this is true in particular for the variable unemployment reported in Tables 3 and 4; it is important to note that these predictors have not been considered for Group 3 since this group includes only LLMA = 403 (Rome);
- comparing the SSD estimator and the composite estimator $\hat{Y}^{C2}$, both combining a direct component with a synthetic component with sample weights, the SSD estimator seems preferable since the perfor mance of the two estimators is very close but SSD is superior in terms of computational simplicity. Since in actual surveys the optimal weights are not known, the present analysis suggests using the SSD estimator; a drawback is that a specific study has to be carried out for the choice of the parameter λ.

Table 3
Percentage Average Absolute Relative Bias and Percentage Average Root Relative Means Square Error of the Estimators of Unemployed by Group of Local Labour Market Areas

Estimator	AARB	ARRMSE Group 1	AARB	ARRMSE Group 2	AARB	ARRMSE Group 3
Expansion	3.52	123.30	0.71	35.01	0.11	6.19
Post-stratified ratio	36.94	72.07	0.77	28.43	0.08	5.68
Synthetic	17.06	18.24	22.68	24.28	5.84	7.33
Composite C1	16.52	17.85	14.71	17.66	2.19	5.50
Composite C2	9.95	35.59	6.89	23.86	3.98	6.68
SSD	10.11	34.77	11.27	19.89	2.99	5.70
EBLUP EB1	13.84	80.14	12.06	29.75	*	*
EBLUP EB2	14.44	91.89	9.02	28.74	*	*

5.2 Conditional Analysis

For the conditional measures we limit ourselves to the presentation of the results for the following four LLMAs: Bagnoregio (code number = 391) and Civita Castellana (code number = 392) in the small group, Cassino (code number = 413) in the medium group, and Rome (code number = 403) for the large group. The frequency distributions over the 2,000 replications of the PSUs' counts in each selected area are very different as a consequence of the LLMAs' different sizes.

Recall that we could not consider EBLUPs for LLMA 403 since it is the only one in GROUP 3.

In Table 4 the results of the study areas are reported for the variable number of Unemployed.

The following points arise:

- the post-stratified ratio estimator usually has conditional bias near zero when the sample size, n_d, takes an inner value of its frequency distribution;
- the post-stratified ratio estimator usually shows better conditional performance, in terms of conditional bias and of RRCMSE, than the expansion estimator;

Table 4
Percentage Relative Conditional Bias and Percentage Relative Root Conditional MSE of the Estimators Conditioned on the Number of Sampled Municipalities for Given LLMAs

Number of sampled Municipalities	Proportion of simulations %	Expansion	Post stratified Ratio	Synthetic	Composite C1	Composite C2	Sample Size Dependent	EBLUP EB1	EBLUP EB2
				LLMA = 391					
				Percentage Relative Conditional Bias					
0	72.43	−100.00	−100.00	28.69	28.69	28.69	28.69	26.76	39.68
1	25.29	208.30	−4.28	28.39	28.21	7.28	−4.28	−2.35	−4.21
2	2.24	527.40	0.66	29.81	29.65	9.22	0.66	1.21	0.97
3	0.05	637.88	−16.53	24.68	24.45	1.41	−16.53	85.54	−12.22
				LLMA = 391					
				Percentage Relative Root Conditional MSE					
0	72.43	100.00	100.00	29.33	29.33	29.33	29.33	141.08	163.51
1	25.29	281.18	68.54	29.03	28.85	48.20	68.54	70.44	66.58
2	2.24	588.02	45.51	30.23	30.07	33.49	45.51	84.47	47.33
3	0.05	637.88	16.53	24.68	24.45	1.41	16.53	85.54	12.22
				LLMA = 392					
				Percentage Relative Conditional Bias					
0	8.79	−100.00	−100.00	10.26	10.26	10.26	10.26	−6.80	−7.45
1	27.32	−48.19	1.51	9.94	9.85	5.31	8.19	−5.31	0.80
2	34.03	−2.01	−3.07	10.30	10.15	1.71	3.85	−5.88	−4.18
3	20.57	43.54	−2.95	10.22	10.08	1.01	0.50	−3.84	−3.64
4	7.65	108.22	4.05	10.88	10.81	6.58	4.34	−6.07	0.98
5	1.39	159.44	3.80	13.33	13.22	6.04	3.80	−6.21	1.06
6	0.25	169.30	−13.82	10.14	9.87	−5.39	−13.82	−14.92	−13.08
				LLMA = 392					
				Percentage Relative Root Conditional MSE					
0	8.79	100.00	100.00	11.47	11.47	11.47	11.47	40.38	43.91
1	27.32	60.11	74.67	11.24	11.19	58.36	21.25	33.99	65.82
2	34.03	48.50	48.03	11.50	11.37	34.87	24.39	28.57	41.19
3	20.57	70.07	38.01	11.54	11.41	27.09	27.52	28.41	32.86
4	7.65	129.85	35.12	11.92	11.87	24.80	33.96	29.47	30.96
5	1.39	171.38	26.29	14.09	13.97	18.80	26.29	26.84	24.56
6	0.25	173.01	20.23	11.07	10.84	15.04	20.23	19.06	17.92
				LLMA = 413					
				Percentage Relative Conditional Bias					
0	0.05	−100.00	−100.00	2.47	2.47	2.47	2.47	−100.00	−100.00
1	1.29	−74.42	8.04	5.60	5.63	8.36	6.08	−9.08	4.88
2	7.40	−49.73	0.92	4.56	4.52	2.72	3.68	−16.72	−2.08
3	21.31	−26.46	0.93	5.06	5.01	2.37	3.55	−15.33	−1.86
4	28.96	−4.60	−1.01	5.11	5.04	1.14	2.26	−17.29	−3.83
5	25.48	19.41	−0.31	4.92	4.86	1.72	1.78	−16.93	−3.18
6	11.43	42.48	0.14	4.64	4.58	1.91	1.45	−16.70	−2.81
7	3.68	66.82	0.86	5.04	4.99	1.77	1.58	−15.11	−1.91
8	0.40	59.75	−14.54	4.74	4.51	−7.72	−13.37	−28.08	−17.24
				LLMA = 413					
				Percentage Relative Root Conditional MSE					
0	0.05	100.00	100.00	2.47	2.47	2.47	2.47	100.00	100.00
1	1.29	76.71	77.02	8.00	8.14	66.06	13.49	72.44	75.79
2	7.40	54.07	46.04	6.69	6.69	36.83	12.75	46.44	45.42
3	21.31	36.86	36.07	7.11	7.09	27.54	14.15	39.35	35.63
4	28.96	32.02	32.26	7.28	7.24	23.93	16.22	36.92	32.12
5	25.48	38.45	27.51	6.97	6.94	20.05	16.99	33.58	27.43
6	11.43	53.61	22.02	6.52	6.47	16.06	15.95	29.75	21.94
7	3.68	77.79	24.58	6.83	6.79	17.86	20.53	28.44	23.81
8	0.40	65.42	18.76	8.19	7.96	12.71	17.42	34.65	21.61
				LLMA = 403					
				Percentage Relative Conditional Bias					
8	0.15	−5.20	3.17	−3.96	0.56	−1.82	−0.71	*	*
9	0.20	−2.87	3.38	−2.10	1.37	−0.67	0.43	*	*
10	1.59	−4.66	−0.15	−5.82	−2.23	−3.45	−3.15	*	*
11	4.82	−2.98	0.36	−6.13	−2.02	−3.53	−3.04	*	*
12	11.38	−2.41	−0.03	−5.98	−2.21	−3.91	−3.11	*	*
13	20.32	−1.52	−0.30	−6.15	−2.44	−4.16	−3.30	*	*
14	23.40	−0.15	−0.10	−5.84	−2.20	−4.05	−3.00	*	*
15	18.68	1.01	−0.07	−5.51	−2.06	−3.93	−2.79	*	*
16	12.67	2.51	0.20	−5.64	−1.94	−3.85	−2.69	*	
17	4.42	3.73	0.25	−5.49	−1.85	−3.66	−2.55	*	
18	1.84	1.86	−2.55	−7.20	−4.25	−6.32	−4.80	*	
19	0.55	6.28	0.71	−4.70	−1.28	−3.24	−1.88	*	

Table 4 (continued)
Percentage Relative Conditional Bias and Percentage Relative Root Conditional MSE of the Estimators Conditioned on the Number of Sampled Municipalities for Given LLMAs

Number of sampled Municipalities	Proportion of simulations %	Expansion	Post stratified Ratio	Synthetic	Composite C1	Composite C2	Sample Size Dependent	EBLUP EB1	EBLUP EB2
				LLMA = 403					
				Percentage Relative Root Conditional MSE					
8	0.15	6.79	5.24	5.52	4.05	4.46	4.04	*	*
9	0.20	5.64	6.06	4.79	4.81	4.69	4.49	*	*
10	1.59	7.81	6.19	6.86	5.49	6.26	5.54	*	*
11	4.82	6.54	5.75	7.51	5.41	6.54	5.66	*	*
12	11.38	6.14	5.56	7.34	5.37	6.61	5.62	*	*
13	20.32	6.34	6.01	7.72	5.86	7.12	6.09	*	*
14	23.40	5.83	5.62	7.23	5.43	6.58	5.63	*	*
15	18.68	5.98	5.58	7.10	5.42	6.51	5.59	*	*
16	12.67	6.02	5.20	7.10	5.07	6.31	5.28	*	*
17	4.42	7.33	5.82	7.16	5.53	6.72	5.66	*	*
18	1.84	6.40	6.38	8.76	6.90	8.56	7.16	*	*
19	0.55	8.38	5.42	6.53	5.04	5.84	5.12	*	*

- synthetic estimators and the composite estimator $\hat{Y}_d^{C1}$ show the best performances in terms of RRCMSE for LLMAs 391, 392, 413 and 403, confirming what was observed in the unconditional analysis. The only relevant exception is for LLMA 403 for the variable Employed (not reported here) where the post-stratified ratio is the best. In fact the variances of the different estimators are very low for this small area so that the bias is decisive;
- in terms of RRCMSE neither $\hat{Y}_d^{C2}$ nor SSD seems to outperform the other.

We have not reported here the results for the conditional coverage rate (CCR), but we can summarize them as follows:

- the post stratified estimator, the composite estimator $\hat{Y}_d^{C2}$ and the SSD estimator have CCR close to the nominal value apart from extremes values of the PSU counts;
- the EBLUPs' CCRs are also close to the nominal value but we suspect this is due to their high variances;
- for the LLMA = 403 and the Employed variable, the CCR of all the estimators is far from the nominal value.

5.3 Conclusions

As we have already observed, the results for the EBLUP estimators are unsatisfactory; the model used is not adequate, likely because we are far from the hypothesis of unbiasedness for the direct component (post-stratified ratio estimator) in many cases and because of the choice of the auxiliary variables. One of the main points we intend to address in future work is the improvement of the explicit models for EBLUP.

The composite estimator $\hat{Y}_d^{C1}$ turns out to be the best in terms of ARRMSE and RRCMSE. If weights are thought to be stable they may be evaluated, for example, at a Census point and $\hat{Y}_d^{C1}$ applied. If sample dependent weights are to be used, then the SSD estimator seems preferable to the composite estimator $\hat{Y}_d^{C2}$ because of its computational simplicity, even if some *ad hoc* study may be necessary for the choice of the parameter λ, since the two estimators' unconditional and conditional properties do not differ greatly. In any case, some improvements can be gained for the composite and SSD estimators through use of a better synthetic estimator, in terms of the number and the choice of post-strata, or in terms of a better choice of the auxiliary variables as observed for the EBLUP.

In this work we have examined conditional and unconditional properties of some common estimators; our interest in the future will be to examine also the empirical properties from the conditional point of view of the conditional estimators proposed in the work by Falorsi and Russo (1999).

References

Cochran, W.G. (1977). *Sampling Techniques*, New York: John Wiley & Sons, Inc.

Falorsi, P.D., and Russo, A. (1996). A conditional analysis of some small area estimators in two-stage sampling. In *Proceedings of 1996 Annual Research Conference*, Bureau of the Census, Washington. 613- 637.

Falorsi, P.D., and Russo, A. (1999). A conditional analysis of some small area estimators in two-stage sampling. *Journal of Official Statistics*, 15, 4, 537-550.

Falorsi, P.D., Falorsi, S. and Russo, A. (2000). A conditional analysis of some small area estimators in sampling with two primary units selected in each stratum. *Statistics in Transitions, Journal of the Polish Statistical Association*, 4, 4, 565-585.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 366, 269-277.

Ghosh, 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*, A, 142, 33-46.

Pfeffermann, D. (1999). Small area estimation – big developments. In *Proceedings of the IASS Satellite Conference Small Area Estimation*, Riga 1999, 129-145.

Purcell, N.J., and Kish, L. (1979). Estimation for small domain. *Biometrics*, 35, 365-384.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.

Rao, J.N.K. (1999). Some recent advances in model based small area estimation. *Survey Methodology*, 25, 175-186.

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.

Russo, A., and Falorsi, P.D. (1993). Conditional and unconditional properties of small area estimators in two-stage sampling. In *Proceedings of the International Scientific Conference of the International Association of Survey Statistician*, Warsaw, 1992. 251-270.

Särndal, C.-E., and Hidiroglou, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

Schaible, W.L. (1978). Choosing weights for composite estimators for small area statistics. In *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.