

VARIANCE ESTIMATION OF SMALL AREA ESTIMATORS IN THE SPANISH LABOUR FORCE SURVEY

Francisco Hernández¹, Montserrat Herrador¹, Domingo Morales², M. Dolores Esteban² y Agustín Pérez².

¹ Instituto Nacional de Estadística

² Centro de Investigación Operativa, Universidad Miguel Hernández de Elche

Abstract

The main goal of this work is to investigate how to estimate sampling design variances of model-based and model-assisted small area estimators in a complex survey sampling setup. For this sake, the Spanish Labour Force Survey is considered. Sample and aggregated data are taken from the Canary Islands in the second trimester of 2003 in order to obtain some small area estimators of ILO unemployment totals and rates. Several problems arising from the application of standard small area estimation procedures to the survey are described. It is shown that standard variance estimators based on explicit formulas are not applicable in strict sense, since the assumptions under which they are derived do not hold. In addition two resampling techniques, bootstrap and jackknife, are considered. These methods treat all the considered estimators in the same manner and therefore they can be used as performance measures to compare them. From the analysis of the obtained results, some recommendations are given.

Key words and phrases: Labour Force Survey, small area estimation, mean square error, bootstrap, jackknife.

AMS subject classification: 62D05, 62J05.

1. Introduction

Almost all the methodological development to date in small area estimation has been carried out under the assumption that the assumed small area model is true; and that the appropriate measure of accuracy of the small area estimator is its repeated sampling variability under random realisations of the population assuming the small area model holds. The fact that the assumed model only approximates reality, and that the measures that capture sampling variability relative to the actual population values are often of primary interest, is often ignored. This paper focuses this imbalance by dealing with the repeated sampling properties of the most commonly used model-based methods of small area estimation.

The paper investigates problems arising from the application of standard small area estimation techniques to complex sampling designs, including nonresponse, reliability of population sizes, selection of auxiliary variables with the same definition in the survey and in the administrative registers, consistency with the officially published data at a higher level of aggregation and estimation of mean squared errors in a complex setup. For this sake, some model-assisted and model-based estimators are adapted to the Spanish Labour Force Survey (SLFS) in order to estimate totals and rates of unemployed people by sex and small areas in the Canary Islands.

In Section 2 some technical details of SLFS are described, with special emphasis on the sampling design, the separated ratio estimator of totals and the calibration of sampling weights. In Section 3 considered small area estimators and explicit formulas to estimate their variances or mean squared errors are introduced. In Section 4 some comments on the selection of auxiliary variables are presented. In Section 5 two-stage bootstrap and jackknife methods are proposed to estimate the sampling variances of the considered small area estimators. These resampling methods produce performance measures to compare estimators of totals and rates. In Section 6

conclusions on the considered small area estimators and on the three introduced methods to estimate their variance are given.

2. The Spanish Labour Force Survey (SLFS)

This quarterly survey follows a stratified two-stage random sampling design with separate samples s_p drawn from each province P_p . The Primary Sampling Units (PSUs) are Census Sections (geographical areas with a maximum of 500 dwellings – approximately 3.000 people) and they are grouping in strata according to the size of municipality. Within each stratum, PSUs are selected with probabilities proportional to size according to the number of dwellings. In the second stage sampling, the Secondary Sampling Units (SSUs) are dwellings and a random start systematic sampling is applied to draw a fixed number (18 in most cases) of SSUs from each selected PSU. All peopled aged 16 years old or more in the selected SSUs are interviewed.

The probability that a dwelling v belonging to PSU a of stratum h be selected in s_p is

$$P(Dwe_{hav}) = P(PSU_{ha})P(Dwe_{hav} | PSU_{ha}) = m_h \frac{V_{ha}}{V_h} \frac{18}{V_{ha}} = \frac{18m_h}{V_h},$$

where V_{ha} and V_h are the totals of dwellings in PSU a of stratum h and in stratum h respectively and m_h is the number of sections allocated in stratum h . Because all individuals in a selected dwelling are interviewed, the inclusion probabilities of individuals and dwellings coincide. Therefore, the inclusion probability of individual j in dwelling v and stratum h is

$$\pi_j = \frac{18m_h}{V_h} = \pi_h,$$

so, given a stratum, all individuals have the same selection probability, i.e. this survey uses what is called a self-weighting design. Afterwards, at stratum level, probabilities π_j are modified to take non-response into account and their inversions produce sampling weights $w_j^{(1)}$ adjusted by non-response. Consequently the survey is still using a self-weighting design inside of each stratum.

Up until year 2001 the INE used a ratio estimator, with Demographic Population Projections as auxiliary variable, to estimate the total Y_p of variable y in the province p , i.e.

$$\hat{Y}_p^{Jfs*} = \sum_{h \in P_p} \frac{N_h}{\hat{N}_h} \sum_{v \in s_h} \sum_{j \in v} w_j^{(1)} y_j \quad \text{with} \quad \hat{N}_h = \sum_{v \in s_h} \sum_{j \in v} w_j^{(1)} = w_j^{(1)} n_h,$$

where N_h is the projection of the population living in familiar dwellings in stratum h , with reference to half the quarter and n_h is the number of people living in the dwellings in the sample, in stratum h , at the time of the interview. Alternatively,

$$\hat{Y}_p^{Jfs*} = \sum_{h \in P_p} \sum_{j \in s_h} \frac{N_h w_h^{(1)}}{\hat{N}_h} y_j = \sum_{j \in s_p} w_j^{(2)} y_j,$$

with the sample dependent weights

$$w_j^{(2)} = w_j^{(2)}(s_p) = \frac{N_h w_h^{(1)}}{\hat{N}_h} = \frac{N_h}{n_h} \quad \text{if} \quad j \in s_h.$$

Since the first quarter of 2002, reweighting (or calibration) techniques are applied to estimators so as to adjust the survey estimates to some given information from external sources. The reweighting technique (see Deville and Särndal (1992)) requires the availability of K auxiliary variables appearing in the sample s_p and whose populations totals are known, i.e.

$$\sum_{j \in P_p} x_{jk} = X_k, \quad k=1, \dots, K.$$

The target is to find a new estimator

$$\hat{Y}_p^{Jfs} = \sum_{j \in s_p} w_j y_j$$

with new weights w_j satisfying the balance equation

$$\sum_{j \in s_p} w_j x_{jk} = X_k$$

and being as similar as possible to $w_j^{(2)}$. The problem aims to find values w_j minimising

$$\sum_{j \in s_p} w_j^{(2)} G(w_j / w_j^{(2)}) \quad \text{restricted to} \quad \sum_{j \in s_p} w_j x_{jk} = X_k, \quad k=1, \dots, K,$$

where G is a function of distance. The solution of the problem depends on G . If the linear distance function, with argument $z = w_j / w_j^{(2)}$, is considered, i.e.

$$G(z) = \frac{1}{2}(z-1)^2, \quad z \in \mathbb{R}$$

then the problem can be solved explicitly by using Lagrange multipliers which facilitate obtaining a set of factors w_j verifying the balance conditions and provide the same estimates as the generalised regression estimator. In the second trimester of 2003 the SLFS weights were calibrated so that their sum coincide with the population projections for individuals aged 16 years and over per groups of sex and age in autonomous communities, and per provinces. In order to obtain the practical solution for this problem, it was employed the CALMAR (CALage sur MARGes) software, programmed in SAS code by the INSEE (Institut National de la Statistique et des Études Économiques) in France.

SLFS estimator of the total Y_p of variable y in province p is \hat{Y}_p^{lfs} . SLFS estimators of the total and the mean of domain d are

$$\hat{Y}_d^{lfs} = \sum_{j \in s_d} w_{dj} y_{dj} \quad \text{and} \quad \hat{\bar{Y}}_d^{lfs} = \frac{\hat{Y}_d^{lfs}}{\hat{N}_d}, \quad \text{with} \quad \hat{N}_d = \sum_{j \in s_d} w_{dj}.$$

For provinces $P_p = \bigcup_{d=1}^{D_p} P_{pd}$, it holds $\hat{Y}_p^{lfs} = \sum_{d \in P_p} \hat{Y}_d^{lfs}$; i.e. there exist consistency between SLFS

estimates at domain and province level.

SLFS publishes estimates of employment and unemployment totals at province level. If in the near future these publications were extended to domain levels it should be necessary to force consistency between both types of data. This is to say that the sum of the estimated totals in all the domains within a province should coincide with the actual estimated total by SLFS in the province. In order to fulfil this consistency criterion the following modification of the small area estimates has been implemented.

Let \hat{Y}_p^{lfs} be the SLFS estimator of total Y_p in province p . Assume that province p is partitioned in D_p domains; i.e. $P_p = \bigcup_{d=1}^{D_p} P_d$ with $P_{d_1} \cap P_{d_2} = \emptyset$ if $d_1 \neq d_2$. Let $\hat{Y}_1, \dots, \hat{Y}_{D_p}$ be some given estimators of totals Y_1, \dots, Y_{D_p} . In general, the consistency property

$$\hat{Y}_p^{lfs} = \sum_{d=1}^{D_p} \hat{Y}_d$$

is not satisfied. In such cases $\hat{Y}_1, \dots, \hat{Y}_{D_p}$ can be transformed into consistent estimators by the following calculation

$$\hat{Y}_d^c = \lambda_{yp} \hat{Y}_d, \quad \text{where} \quad \lambda_{yp} = \frac{\hat{Y}_p^{lfs}}{\sum_{d=1}^{D_p} \hat{Y}_d}$$

are *consistency factors*. For consistent estimators, it holds

$$\hat{Y}_p^{lfs} = \sum_{d=1}^{D_p} \hat{Y}_d^c.$$

3. Estimators of small areas totals in the SLFS

In this section we introduce some estimators of the total a target variable y . Unemployment rate estimators are obtained by substituting in its formula totals by its corresponding estimators. Explicit formulas to estimates sampling variance of model-assisted estimators and to estimate mean squared errors of model-based estimators are also given.

3.1. Direct estimator

In this work *direct* estimator is the one appearing in Särndal et al. (1992), p. 391, when N_d is *known*, but substituting theoretical weights by calibrated ones. Its expression is

$$\hat{Y}_d^{direct} = N_d \hat{Y}_d^{lfs}.$$

An explicit-formula estimator of its sampling variance is

$$\text{var}(\hat{Y}_d^{direct}) = \left(\frac{N_d}{\hat{N}_d} \right)^2 \sum_{j \in s_d} w_j (w_j - 1) (y_j - \hat{Y}_d^{direct})^2$$

3.2. GREG estimator

Consider p explanatory variables measured at N population units; i.e. $\mathbf{x}_j = (x_{j,1}, \dots, x_{j,p})$, $j=1, \dots, N$. Let

$$\bar{\mathbf{X}}_d = \frac{1}{N_d} \sum_{j \in P_d} \mathbf{x}_d \quad \text{and} \quad \hat{\bar{\mathbf{X}}}_d^{lfs} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_j \mathbf{x}_j$$

be population and direct means. Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where \mathbf{X} is an $n \times p$ matrix with rows \mathbf{x}_j , $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{W}^{-1})$ and $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$. The weighted least square estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W} \mathbf{y} = \left(\sum_{j \in s} w_j \mathbf{x}_j^t \mathbf{x}_j \right)^{-1} \left(\sum_{j \in s} w_j \mathbf{x}_j^t y_j \right).$$

Observe that in the set of p explanatory variables one can include artificial variables. In this work first variable is such that $x_{j,1} = 1$, $j=1, \dots, n$; i.e. we assume a linear model with intercept term. In this way, estimation of $\boldsymbol{\beta}$ does not depend on the type of selected small area in territories with hierarchical structure.

GREG estimator of a total (see e.g. Särndal et al., 1992, p. 410) is

$$\hat{Y}_d^{greg} = N_d \hat{Y}_d^{lfs} + N_d (\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d^{lfs}) \hat{\boldsymbol{\beta}}.$$

Observe that

$$\hat{Y}_d^{greg} = \sum_{j \in s} g_{dj} w_j y_j \quad \text{and} \quad N_d \bar{\mathbf{X}}_d = \sum_{j \in s} g_{dj} w_j \mathbf{x}_j,$$

where

$$g_{dj} = \frac{N_d}{\hat{N}_d} I_{P_d}(j) + N_d (\bar{\mathbf{X}}_d - \hat{\bar{\mathbf{X}}}_d^{lfs}) \left(\sum_{j \in s} w_j \mathbf{x}_j^t \mathbf{x}_j \right)^{-1} \mathbf{x}_j^t.$$

An explicit-formula estimator of its sampling variance is

$$\text{var}(\hat{Y}_d^{greg}) = \sum_{j \in s_d} w_j (w_j - 1) g_{dj}^2 (y_j - \mathbf{x}_j \hat{\boldsymbol{\beta}})^2.$$

3.3. EBLUPA estimator

EBLUPA estimator is a composite estimator based on the 2-level linear mixed model (model A)

$$y_{dj} = \mathbf{x}'_{dj}\boldsymbol{\beta} + u_d + v_{dj}^{-1/2}e_{dj}, \text{ where } u_d \sim iid N(0, \sigma_u^2) \text{ and } e_{dj} \sim iid N(0, \sigma_e^2) \text{ are independent.}$$

The model is fitted by calculating maximum likelihood estimators of the regression and variance component parameters with a Fisher-scoring algorithm (see e.g. Rao, 2003, ch. 5-6).

EBLUPA estimator of a total is $\hat{Y}_d^{eblupa} = N_d \hat{Y}_d^{eblupa}$, where

$$\hat{Y}_d^{eblupa} = \hat{\gamma}_d (\hat{Y}_d^{lfs} - \hat{\mathbf{X}}_d^{lfs} \hat{\boldsymbol{\beta}}) + \bar{\mathbf{X}}_d \hat{\boldsymbol{\beta}}, \text{ with } \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\hat{\sigma}_e^2 / v_d)}, \quad v_d = \sum_{j \in s_d} v_{dj}.$$

EBLUPA estimator is in fact a pseudo-eblup estimator studied in work package 4 of EURAREA project (<http://www.statistics.gov.uk/eurarea/>) and related to the ones proposed by Prasad and Rao (1999) and You and Rao (2002). Mean squared error is estimated by using g₁-g₄ explicit formulas given by Prasad and Rao (1990) and later extended by Das, Jiang and Rao (2001) to more general linear mixed models.

3.4. EBLUPB estimator

EBLUPB estimator is a composite estimator based on the area-level model (model B)

$$\bar{Y}_d = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d \quad \text{and} \quad \hat{Y}_d^{lfs} = \bar{Y}_d + \varepsilon_d,$$

where $u_d \sim iid N(0, \sigma_u^2)$ and $\varepsilon_d \sim iid N(0, \sigma_d^2)$ are independent. This model was introduced by Fay and Herriot (1979) to estimate average per capita income for small areas in USA. The model is fitted by the same method as model A. Under model B, EBLUP estimator of total is

$$\hat{Y}_d^{eblupb} = \hat{\gamma}_d \hat{Y}_d^{lfs} + (1 - \hat{\gamma}_d) \bar{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}}, \quad \text{with} \quad \hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_d^2}$$

Mean squared error is estimated by using g₁-g₃ explicit formulas given by Prasad y Rao (1990).

4. Selection of auxiliary variables in the SLFS

To obtain models with high predictive properties, the selection of adequate explanatory variables is very important. In the case of individual level models, auxiliary variables are needed at both individual and domain level. At the individual level auxiliary variables are obtained from the survey sample and, except for the cases of nonresponse, their values are available. However it is much more difficult to evaluate auxiliary variables at the domain level, because their values come from external sources which sometimes are not available, have not sufficiently good quality or may even present definition differences with their sample counterparts. Because of these reasons, the number of available auxiliary variables for individual-level models, describing employment or unemployment, has been very small. In addition it has not been possible to find an adequate continuous variable for these models.

In relation to the application of SAE methods to real data relative to one region of Spain, some technical details are listed in next lines:

1. *Universe* of interest is the region of Canary Islands. To fit models the entire sample data of the universe is jointly used.
2. *Domains* of interest are small areas (provisional geographical divisions for statistical purposes) crossed with sex. There are 2×27=54 domains in the considered universe.
3. An *auxiliary variable aggregated at the domain level* without sample counterpart has been used. This variable is named GSAU and consists in 6 groups of sex – age classified as unemployed in the administrative register of employment claimants with 12 values at all.
4. The following *auxiliary variables* have been used at aggregated and sample level:

- GSA: groups of sex – age with 6 values. Three age groups have been considered: 16-24, 25-54 and ≥ 55 .
 - GSAC: groups of sex - age – employment claimant with 12 values,
 - BIPSTRATUM: groups of province – bistratum with 4 values.
- Variable BISTRATUM, used in the definition of BIPSTRATUM, is calculated as follows:
- BISTRATUM = 1 if SLFS stratum is 1, 2, 3, 4,
 - BISTRATUM = 2 otherwise.
5. Estimation procedures for variable UNEMPLOYED have used the following auxiliary variables:
 - BIPSTRATUM and GSAC for estimators GREG and EBLUPA.
 - BIPSTRATUM and GSAU for estimators EBLUPB.
 6. Estimation procedures for variable EMPLOYED have used following auxiliary variables:
 - BIPSTRATUM and GSA for estimators EBLUPA and EBLUPB.
 - BIPSTRATUM and GSAC for estimator GREG.

Among the considered (although not finally used) auxiliary variables, PSTRATUM (groups of province - stratum) is the one with highest explanatory power since, in SLFS survey, strata are indicators of degree of rurality, with the value 1 (minimum degree) at the capital of the province and the value 9 (maximum degree) at the municipalities with less than 2000 inhabitants. PSTRATUM was used as explanatory variable in GREG models (standard linear models with intercept and without fixed effects in domains) and in EBLUPA models (linear mixed models with random effects in domains) for dependent variables UNEMPLOYED and EMPLOYED, and also in the construction of the derived small area estimators of totals and rates of unemployed people. PSTRATUM is significant in such models but parameters associated to its categories are estimated with high variances, resulting non stable fittings. Note that for models with PSTRATUM and GSAC as explanatory variables, $12+12=24$ parameters have to be estimated in the Canary Islands.

There exists a second argument against the use of PSTRATUM as auxiliary variable in the construction of GREG and EBLUPA estimators. To calculate such estimators one needs aggregated data in all the intersections of domains with categories of the auxiliary variable. So reliable population sizes are needed in the intersections of sex, stratum and small areas, but if the quality of the population sizes is not good enough, it is better to reduce the explanatory power of the auxiliary variable and to substitute PSTRATUM by BIPSTRATUM in order to improve the quality of the aggregated external data.

Tables 4.1-4.2 present ANOVA analysis for GREG models to show that all considered explanatory variables are significant.

	d.f.	Sum of squares	Mean of squares	F	Pr(>F)
BIPSTRATUM	13	5,64	0,43	14,854	< 2,2e-16
GSAC	11	480,48	43,68	1494,452	< 2,2e-16
Residual	26161	764,64	0,03		

Table 4.1. ANOVA for variable UNEMPLOYED.

	d.f.	Sum of squares	Mean of squares	F	Pr(>F)
BIPSTRATUM	13	39,1	3,0	17,832	< 2,2 e-16
GSA	5	2090,4	418,1	2476,962	< 2,2 e-16
Residual	26167	4416,7	0,2		

Table 4.2. ANOVA for variable EMPLOYED.

5. Resampling methods for design-based variance estimation in the SLFS

In this section we describe a two-stage bootstrap method as well as a two-stage jackknife method to estimate variances of small area estimators of totals and rates.

5.1. Two-stage bootstrap method

Let θ be a parameter to be estimated with $\hat{\theta}$. Bootstrap (see e.g. Efron and Tibshirani, 1998) is a resampling method which is often used to estimate variances $Var(\hat{\theta})$. To implement the proposed two-stage bootstrap method, it is not necessary to construct artificial populations since the procedure generates bootstrap samples directly from the original SLFS sample as it is explained in next lines.

Let s be an SLFS sample in a given province. Let $s = \bigcup_{h=1}^H s_h$, where s_1, \dots, s_H are subsamples by strata. Let $s_h = \bigcup_{a=1}^{m_h} s_{ha}$, where s_{h1}, \dots, s_{hm_h} are subsamples in the m_h selected PSUs from the stratum h . Finally, let $s_{ha} = \bigcup_{v=1}^{m_{ha}} s_{hav}$, where $s_{ha1}, \dots, s_{ham_{ha}}$ are the subsamples in the m_{ha} visited dwellings in PSU a from stratum h . Selection of bootstrap samples in stratum $h=1, \dots, H$, is done in the following way:

1. Select a simple random sample with replacement of m_h PSUs from the set of m_h PSUs appearing in the original SLFS sample.
2. Within each selected PSU, draw a simple random sample with replacement of m_{ha} dwellings from the set of m_{ha} dwellings appearing in the given PSU of the original SLFS sample.
3. Select all the individuals aged 16 or more from the dwellings in the bootstrap sample.

Variance estimation is done as follows:

- A. By using the procedure described above, use sample s to draw B bootstrap samples. For every bootstrap sample calculate $\hat{\theta}_b^*$, $b=1, \dots, B$, in the same way as $\hat{\theta}$ was calculated. So, in each bootstrap sample, the weights $w_{j,b}^{*(2)} = N_h / n_{hb}^*$ (where n_{hb}^* is the number of individuals selected in bootstrap sample b and stratum h) are adjusted by a calibration procedure to obtain calibration weights w_j^* in the same way as in SLFS sample (see Section 2). These calibration weights w_j^* are used to calculate $\hat{\theta}_b^*$.
- B. The observed distributions of $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ is expected to imitate the distribution of estimator $\hat{\theta}$ in the SLFS sampling design.
- C. The variance of $\hat{\theta}$ is approximated by

$$\text{var}_B(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}^*)^2, \quad \text{where} \quad \hat{\theta}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*.$$

- D. A bootstrap estimator of the sampling error (*coefficient of variation*) in % of $\hat{\theta}$ is

$$cv_B(\hat{\theta}) = \frac{\sqrt{\text{var}_B(\hat{\theta})}}{\hat{\theta}} \cdot 100.$$

An important step when estimating variances through the bootstrap method is to take into account the consistency property of estimators of totals at province level. Consistency property was not required in the bootstrap samples. To estimate variances of consistent estimators, estimated variances of non consistent estimators are multiplied by the square of the consistency factor λ (cf. Section 2). However, for the coefficient of variation this adjustment is not necessary. More concretely, if $\hat{\theta}^c = \lambda \hat{\theta}$ is the consistent version of a total estimator $\hat{\theta}$, where λ is the consistency factor calculated in the original SLFS sample, then bootstrap estimators of the variance and the coefficient of variations of $\hat{\theta}^c$ are

$$\text{var}_B(\hat{\theta}^c) = \lambda^2 \text{var}_B(\hat{\theta}) \quad \text{and} \quad cv_B(\hat{\theta}^c) = cv_B(\hat{\theta}).$$

5.2. Delete-one-cluster jackknife method

In order to apply the jackknife for variance estimation in SLFS samples, we use the delete-one-cluster jackknife method (see e.g. Rao and Tausi, 2004). To obtain the delete-one-cluster jackknife variance estimator of $\hat{\theta}$, we generate a jackknife sample by deleting a PSU. So within each province, there are as many jackknife samples as PSU are in the corresponding SLFS sample.

Consider the jackknife simple, $s_{(g,j)}^*$, obtained by excluding PSU j of stratum g . jackknife weight of individual k of PSU i from stratum h in jackknife sample $s_{(g,j)}^*$ is

$$w_{hik(g,j)} = w_{hik}^{(2)} b_{hi(g,j)},$$

where

$$b_{hi(g,j)} = \begin{cases} \frac{m_g}{m_g - 1} & \text{si } h = g, i \neq j, \\ 1 & \text{si } h \neq g, \end{cases}$$

and m_g is the number of PSUs in the stratum g . Note that the case $h=g$ and $i=j$ does not appear in the jackknife sample $s_{(g,j)}^*$. If L is the number of strata in the sample, the variance estimation is done as follows:

- A. By using the procedure described above, use sample s to draw jackknife samples $s_{(g,j)}^*$, $g=1, \dots, L, j=1, \dots, m_g$. For every jackknife sample calculate $\hat{\theta}_{(g,j)}^*$ in the same way as $\hat{\theta}$ was calculated. So, in each jackknife sample, the weights $w_{hik(g,j)}$ are adjusted by a calibration procedure to obtain calibrated weights $w_{hik(g,j)}^*$ in the same way as it was done with the SLFS sample (see Section 2). These calibrated weights $w_{hik(g,j)}^*$ are used to calculate $\hat{\theta}_{(g,j)}^*$.
- B. The observed distributions of $\{\hat{\theta}_{(g,j)}^* : g=1, \dots, L; j=1, \dots, m_g\}$ is expected to imitate the distribution of estimator $\hat{\theta}$ in the SLFS sampling design.
- C. The variance of $\hat{\theta}$ can be approximated by

$$\text{var}_j(\hat{\theta}) = \sum_{g=1}^L \frac{m_g - 1}{m_g} \sum_{j=1}^{m_g} (\hat{\theta}_{(g,j)}^* - \hat{\theta})^2.$$

- D. A bootstrap estimator of the sampling error (*coefficient of variation*) in % of $\hat{\theta}$ is

$$\text{emr}_j(\hat{\theta}) = \frac{\sqrt{\text{var}_j(\hat{\theta})}}{\hat{\theta}} \cdot 100.$$

6. Concluding remarks

If we compare the numerical results obtained with the three considered methods to estimate variances or MSEs, we obtain the following conclusions:

- In domains with large sample size, the three methods produce basically the same results.
- Bootstrap method gives higher estimations of the variances than the explicit-formula or jackknife methods, so it seems that our implementation is positively biased.
- Assumptions required to deriving explicit formulas to estimate variances or MSEs do not hold in practice, so their use should have an orientate character.
- Jackknife method avoids the theoretical problem of the explicit-formula methods and the difficulty of implementation of the bootstrap method. It works quite well in all the domains except in those one with very few sampled PSUs.

Acknowledgements. The authors would like to thank the INE household sampling design unit for their support and helpful comments.

References

- Das, K., Jiang, J. y Rao, J.N.K. (2001). Mean squared error of empirical predictor. *The Annals of Statistics*, **32**, 818-840.
- Deville J.C. and Särndal C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Society*, **87**, 376-382.
- Efron, B. and Tibshirani, R.J. ((1998). *An introduction to the Bootstrap*. Chaoman & Hall/CRC.
- 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**, 269-277.
- Rao, J.N.K. (2003). *Small area estimation*. John Wiley.
- Rao, J.N.K. and Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistic. Theory and methods*. 33, 9, 2087-2095.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Prasad, N. G. N. and Rao, J. N. K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, **85**, 163-171.
- Prasad N.G.N. and Rao J.N.K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*, 25, 67-72.
- You, Y and Rao J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small-area estimation using survey weights. *Survey Methodology*, 30, 431-439.