SORT 28 (2) July-December 2004, 215-230

A comparative study of small area estimators[⋆]

Laureano Santamaría[*], Domingo Morales[*], Isabel Molina[**]

[*] *Universidad Miguel Hernández de Elche*, [**] *Universidad Carlos III de Madrid*

Abstract

It is known that direct-survey estimators of small area parameters, calculated with the data from the given small area, often present large mean squared errors because of small sample sizes in the small areas. Model–based estimators borrow strength from other related areas to avoid this problem. How small should domain sample sizes be to recommend the use of model-based estimators? How robust small area estimators are with respect to the rate *sample size/number of domains*?

To give answers or recommendations about the questions above, a Monte Carlo simulation experiment is carried out. In this simulation study, model-based estimators for small areas are compared with some standard design-based estimators. The simulation study starts with the construction of an artificial population data file, imitating a census file of an Statistical Office. A stratified random design is used to draw samples from the artificial population. Small area estimators of the mean of a continuous variable are calculated for all small areas and compared by using different performance measures. The evolution of this performance measures is studied when increasing the number of small areas, which means to decrease their sizes.

MSC: 62D05, 62F10, 62J02

Keywords: small area estimation, *eblup* estimators, sampling designs, mixed linear models

1 Introduction

The problem of small area estimation arises when samples are drawn from (large) populations, but estimates calculated using sample data are required for smaller domains, within which sample data is not enough to provide reliable direct-survey estimators.

[⋆] Supported by the grants BMF2003-04820 and GV04B-670.

[*] Centro de Investigación Operativa, Universidad Miguel Hernández de Elche. Avda. de la Universidad, s/n, 03202-Elche, Spain.

[**] Departamento de Estadística, Universidad Carlos III de Madrid. C/Madrid, 126, 28903-Getafe, Madrid, Spain.

Received: March 2004
Accepted: July 2004

Usually, any geographical population is partitioned at several levels, which are often nested; for instance, Spain is divided in 19 Communities, where each Community is divided in several provinces, each province in counties, and each county is divided in administrative districts. Thus a design-based estimator may be accurate enough when calculated at the Community or the province levels, but its accuracy may become unacceptable for counties and districts. In that cases, model-based estimators decrease their mean square error by using auxiliary information in the form of regression models. Here we are interested in observing and analyzing the effect of decreasing the level of aggregation (i.e., decrease the rate *total sample size/number of domains*) in the behaviour of some design-based estimators and some model-based estimators, in order to know till which level are design-based estimators reliable and when is necessary to attend to model-based estimators. Further, it is reasonable to suppose that as long as we decrease area sizes, extra "useful" auxiliary information is needed, so models with more information should provide better estimators of small areas. Models with small area random effects are becoming rather popular. Here we compare the performance of fixed effects models and random effects models for the simulated data.

For this purpose, an artificial population is generated imitating a census data file of some geographical population. This file contains the variables defining domains or areas at six nested levels of aggregation, a variable used for stratification, three auxiliary variables, one of them categorical, and the target variable. From this artificial population, 10,000 samples have been extracted. For each sample, small-area estimators have been calculated at each level of aggregation. At the end, two efficiency measures have been computed for each estimator and each level of aggregation: the first measuring the bias, and the second the mean squared error. The evolution of such measures is investigated when decreasing the level of aggregation.

2 The artificial population

The artificial population is a data file with 11 variables and 300,000 records. Each record represents a household of an imaginary country. The file is generated with the purpose of simulating surveys on income and living conditions. See the description of the file in Table 2.1. The first 7 variables are geographical characteristics, where D_1-D_6 have a nested structure and define the domains or areas, while H, representing strata, produces cross-sections with D_1-D_6. The last 4 variables represent household characteristics. For variable G (socioeconomic condition group), we assume that labour activities are classified into two groups: "better paid" and "worse paid", denoting the first group by BPA.

***Table** 2.1: Description of the Artificial Population.*

variable	position	name and description	values
		geographical characteristics	
D_1	1	*Region*	1–8
D_2	2–3	*Community*	1–16
D_3	4–5	*Province*	1–32
D_4	6–7	*County*	1–64
D_5	8–10	*District*	1–128
D_6	11–13	*Zone*	1–256
H	14	*Stratum*	1–6
		household characteristics	
X_1	15–16	*Total number of household members*	01–30
X_2	17–20	*Total area of the dwelling (m^2)*	0000–9999
G	21	*Socioeconomic condition group*	1–4
		All members of the household are unemployed	1
		There are employed members. None of them in BPA	2
		There are employed members, but only one in BPA	3
		There are employed members. Two or more in BPA	4
		target variable	
Y	22–26	*Total net monetary annual income of the household*	00000–99999

Generation of stratum-zone sizes

Let N_{hd_6} be the number of households on stratum h and zone d_6, $h = 1, \dots, 6$, $d_6 = 1, \dots, 256$. These numbers are generated according to the following algorithm

1. Generate $6 \times 256 = 1,536$ uniform numbers in the interval $(0, 1)$. Denote these numbers by u_{hd_6}, $h = 1, \dots, 6$, $d_6 = 1, \dots, 256$.
2. Calculate $u = \sum_{h=1}^{6} \sum_{d_6=1}^{256} u_{hd_6}$ and $v_{hd_6} = u_{hd_6}/u$, $h = 1, \dots, 6$, $d_6 = 1, \dots, 256$.
3. Calculate $N_{hd_6} = [300,000\, v_{hd_6}]$, $h = 1, \dots, 6$, $d_6 = 1, \dots, 256$, where $[\cdot]$ denotes "integer part".
4. If $\sum_{h=1}^{6} \sum_{d_6=1}^{256} N_{hd_6} = 300,000 - n$, with $n > 0$, then add one to the first n sizes N_{hd_6}. Use lexicographic order in subindexes (h, d_6).
5. If $\sum_{h=1}^{6} \sum_{d_6=1}^{256} N_{hd_6} = 300,000$, then stop.

Generation of geographical characteristics

Imputation of numerical values to variables H and D_6 is done by assigning sequentially $H = h$ and $D_6 = d_6$ to N_{hd_6} records, $h = 1, \dots, 6$, $d_6 = 1, \dots, 256$. Variable D_5 is calculated from D_6 by applying formula

$$D_5 = \left[\frac{D_6 + 1}{2} \right].$$

Similar formulas are used to generate D_4, D_3, D_2 and D_1, i.e.

$$D_4 = \left[\frac{D_5 + 1}{2} \right], \quad D_3 = \left[\frac{D_4 + 1}{2} \right], \quad D_2 = \left[\frac{D_3 + 1}{2} \right], \quad D_1 = \left[\frac{D_2 + 1}{2} \right].$$

In this way, the number of areas are doubled from D_ℓ to $D_{\ell+1}$, which means that sample sizes are approximately divided by two.

Generation of household characteristics

- X_1 is generated by

$$\begin{cases} X_1 \sim Poisson(\lambda_{hd_6}) + 1 & \text{if} \quad Poisson(\lambda_{hd_6}) + 1 < 30 \\ X_1 = 30 & \text{otherwise,} \end{cases}$$

  where $\lambda_{hd_6} = 0.8 + 1.5U + h/6 + d_6/256$, $h = 1, \dots, 6$, $d_6 = 1, \dots, 256$ and $U \sim Uniform(0, 1)$.
- X_2 is simulated from

$$X_1 \sim \mathcal{N}(\mu_{hd_6}, \sigma^2_{hd_6}),$$

  where

$$\mu_{hd_6} = 80 + 20U + 2h, \quad \sigma_{hd_6} = 5 + \frac{2d_6}{256}, \quad h = 1, \dots, 6, \; d_6 = 1, \dots, 256$$

  and $U \sim Uniform(0, 1)$.
- G is simulated, conditionally on X_1, from the following discrete distributions:

$$G|_{X_1=1} \sim \begin{pmatrix} 0.1 & 0.5 & 0.4 \\ 1 & 2 & 3 \end{pmatrix} \quad \text{and} \quad G|_{X_1=2,3,\dots} \sim \begin{pmatrix} 0.05 & 0.45 & 0.4 & 0.1 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

  This is to say that for the variable G, two cases are considered. If $X_1 = 1$, G is simulated from a discrete distribution taking values 1, 2, 3 with probabilities 0.1, 0.5 and 0.4 respectively. If $X_1 = 2, 3, \dots$, G is simulated from a discrete distribution taking values 1, 2, 3, 4 with probabilities 0.05, 0.45, 0.4 and 0.1 respectively.

Generation of target variable

Y is simulated from the normal mixed model

$$\begin{aligned} &Y_{hd_6gj} = u_{d_6} + a_h + b_g + \beta_1 X_{hd_6gj1} + \beta_2 X_{hd_6gj2} + e_{hd_6gj}, \\ &h = 1, \dots, 6, \; d_6 = 1, \dots, 256, \; g = 1, \dots, 4, \; j = 1, \dots, N_{hd_6g}, \end{aligned} \tag{2.1}$$

where u_{d_6} and e_{hd_6gj} are the zone and household level residuals, which are independent random variables with distributions $\mathcal{N}(0, \sigma_u^2)$ and $\mathcal{N}(0, \sigma_e^2)$ respectively. Indexes h, d_6, g and j are used to denote stratum, zone, socioeconomic group and household respectively. Therefore, N_{hd_6g} is the number of households in stratum h, zone d_6 and group g, and Y_{hd_6gj}, X_{hd_6gj1}, X_{hd_6gj2} are the values that Y, X_1, X_2 take on the household j of the group g, zone d_6 and stratum h.

The following parameter values are used to generate the artificial population: $\beta_1 = 500$, $\beta_2 = 25$, $\sigma_u^2 = 1000$, $\sigma_e^2 = 750$, $a_h = 4000 + 300h$, $h = 1, \dots, 6$, and $b_g = 5000 + 500g$, $g = 1, \dots, 4$.

Target parameters

Let us use index d_ℓ for geographical characteristic D_ℓ, $\ell = 1, \dots, 6$. Target parameters are

$$\overline{Y}_{d_\ell} = \frac{1}{N_{d_\ell}} \sum_{h=1}^{6} \sum_{g=1}^{4} \sum_{j=1}^{N_{hd_\ell g}} Y_{hd_\ell gj}, \quad \ell = 1, \dots, 6,$$

where $N_{hd_\ell g}$ and $Y_{hd_\ell gj}$ are defined in the same way as $N_{hd_6 g}$ and $Y_{hd_6 gj}$, and N_{d_ℓ} is the number of households in domain $D_\ell = d_\ell$.

3 Notation and estimators

3.1 Notation

The following notation is used

- *Indexes:* s is used for sample and r for nonsample, $d = 1, \dots, D$ for small areas defined by one of the variables $D_1 - D_6$, $g = 1, \dots, G$ for socioeconomic group, and finally $j = 1, \dots, n$ for households.
- *Sizes:* N for population and n for sample. When N or n have indexes they denote size of the corresponding indexed set. For example, n_h is the sample size of stratum h.
- *Totals:* Y or X. When Y or X have indexes they denote the total of the corresponding indexed set. For example, Y_d denotes the total in small area d.
- *Means:* $\overline{Y}$ or $\overline{X}$. When $\overline{Y}$ or $\overline{X}$ have indexes they denote the mean of the corresponding indexed set. For example, $\overline{Y}_d$ denotes the mean of small area d.
- *Weights:* w_j is used for household j and is defined as the inverse of the inclusion probability of household j in the sample. Also, when w has indexes it denotes the sum of weights of the corresponding indexed set.

3.2 Design-based estimators

The following design-based estimators (see e.g. Särndal, Swensson and Wretman (1992)) are considered:

- *w-direct estimator:* It is the classical direct estimator.

$$\widehat{\overline{Y}}_d^{wdirect} = \frac{\sum_{j \in s \cap d} w_j Y_j}{\widehat{N}_d^{direct}}, \quad \widehat{N}_d^{direct} = \sum_{j \in s \cap d} w_j \,.$$

- *Basic synthetic estimator:* It relies on the idea that the population is partitioned in groups larger than areas, for which direct estimators with good precision are available. It is approximately unbiased when all areas contained in a group have the same mean as the whole group. This estimator was used by the USA National

Center for Health Statistics in 1968.

$$\widehat{\overline{Y}}_d^{synt} = \frac{1}{N_d} \sum_{g=1}^{G} N_{dg} \widehat{\overline{Y}}_g^{wdirect} = \frac{1}{N_d} \sum_{g=1}^{G} N_{dg} \left(\frac{\sum_{j \in s \cap g} w_j Y_j}{\sum_{j \in s \cap g} w_j} \right).$$

- *Sample size dependent estimator:* It is constructed by composition of the *w-direct* and the *basic synthetic* estimators. It was proposed by Drew, Sigh and Chouldry (1982).

  $$\widehat{\overline{Y}}_d^{ssd} = \gamma_d \widehat{\overline{Y}}_d^{wdirect} + (1 - \gamma_d) \widehat{\overline{Y}}_d^{synt} ,$$

  where γ_d is calculated by the following formula

  $$\gamma_d = \begin{cases} 1 & \text{if } \widehat{N}_d^{direct} \geq N_d, \\ \frac{\widehat{N}_d^{direct}}{N_d} & \text{otherwise.} \end{cases}$$

3.3 Generalized regression estimators

These estimators arise from fitting a model with general shape

$$\boldsymbol{y} = \boldsymbol{X\beta} + W^{-1/2} \mathbf{e} , \tag{3.1}$$

(see e.g. Section 10.5 of Särndal, Swensson and Wretman (1992)) where $\boldsymbol{y}$ is the vector of sample observations of the target variable Y, $\boldsymbol{X}$ is the matrix whose columns are the observations of auxiliary variables,$\boldsymbol{\beta}$ is the vector of coefficients of mentioned variables, W is a diagonal matrix of known positive elements, and $\boldsymbol{e}$ is the vector of individual errors, satisfying $\boldsymbol{e} \sim N(0, \sigma_e^2 I_n)$, where I_n denotes the identity matrix of size n. Fitting this model by weighted least squares, we get individual predictions

$$\widehat{Y}_{dj} = \boldsymbol{x}_{dj} \widehat{\boldsymbol{\beta}}, \; d = 1, \ldots, D, \; j = 1, \ldots, N_d , \tag{3.2}$$

where $\boldsymbol{x}_{dj}$ represents the row of matrix $\boldsymbol{X}$ corresponding to household j of area d, and

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t W \boldsymbol{X})^{-1} \boldsymbol{X}^t W \boldsymbol{y} . \tag{3.3}$$

Then, the prediction of area mean

$$\overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} Y_{dj}$$

is the mean of predictions of individual values

$$\widehat{\overline{Y}}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \widehat{Y}_{dj} . \tag{3.4}$$

We have used four specific models of the form (3.1), starting from a simple model, and sequentially including more information to the model. Each model provides a generalized regression estimator.

- *Generalized regression synthetic estimator (gsynt):* It is based on the model with common intercept α and X_1 (number of members of the household) as explanatory variable, i.e.

$$y_{dj} = \alpha + \beta x_{dj1} + w_{dj}^{-1/2} e_{dj},\ d = 1, \dots, D,\ j = 1, \dots, n_d. \tag{3.5}$$

This model is a particular case of model (3.2) taking $\beta = (\alpha, \beta)^t$ and $\boldsymbol{x}_{dj} = (1, x_{dj1})$. Applying formula (3.3), estimator $\widehat{\beta} = (\widehat{\alpha}, \widehat{\beta})^t$ is obtained, where $\widehat{\alpha} = \widehat{\overline{Y}}^{wdirect} - \widehat{\beta}\widehat{\overline{X}}_1^{wdirect}$. Replacing $\widehat{\alpha}$ in predictions $\widehat{Y}_{dj} = \widehat{\alpha} + \widehat{\beta} x_{dj1}$, we get

$$\widehat{Y}_{dj} = \widehat{\overline{Y}}^{wdirect} + \widehat{\beta}(x_{dj1} - \widehat{\overline{X}}_1^{wdirect}).$$

Making the average of predictions as in (3.4), we get the following expression of the *gsynt* estimator

$$\widehat{\overline{Y}}_d^{gsynt} = \widehat{\overline{Y}}^{wdirect} + \widehat{\beta}(\overline{X}_{d1} - \widehat{\overline{X}}_1^{wdirect}).$$

- *Generalized regression estimator 1 (greg1):* Here the model is built by replacing the common intercept of model (3.5) by area fixed effects u_d, i.e.

$$y_{dj} = u_d + \beta x_{dj1} + w_{dj}^{-1/2} e_{dj},\ d = 1, \dots, D,\ j = 1, \dots, n_d. \tag{3.6}$$

Here $\beta = (u_1, \dots, u_{d-1}, u_d, u_{d+1}, \dots, u_D, \beta)^t$ and $\boldsymbol{x}_{dj} = (0, \dots, 0, 1, 0, \dots, 0, x_{dj1})$. Again, using formula (3.3), estimators of u_d, $d = 1, \dots, D$ and β are obtained. Replacing formulas of estimators $\widehat{u}_d$, $d = 1, \dots, D$ in individual predictions (3.2), and averaging, we get the following expression of the $greg1$ estimator

$$\widehat{\overline{Y}}_d^{greg1} = \widehat{\overline{Y}}_d^{wdirect} + \widehat{\beta}(\overline{X}_{d1} - \widehat{\overline{X}}_{d1}^{wdirect}).$$

- *Generalized regression estimator 2 (greg2):* In this case the model is obtained from (3.6) by incorporating a second explicative variable, X_2 (total area of the dwelling in m^2), so that

$$y_{dj} = u_d + \beta_1 x_{dj1} + \beta_2 x_{dj2} + w_{dj}^{-1/2} e_{dj},\ d = 1, \dots, D,\ j = 1, \dots, n_d. \tag{3.7}$$

By the same procedure as before, we get the $greg2$ estimator

$$\widehat{\overline{Y}}_d^{greg2} = \widehat{\overline{Y}}_d^{wdirect} + \widehat{\beta}_1(\overline{X}_{d1} - \widehat{\overline{X}}_{d1}^{wdirect}) + \widehat{\beta}_2(\overline{X}_{d2} - \widehat{\overline{X}}_{d2}^{wdirect}),$$

- *Generalized regression estimator 3 (greg3):* This estimator is based on the model

$$y_{dgj} = u_d + b_g + \beta_1 x_{dgj1} + \beta_2 x_{dgj2} + w_{dgj}^{-1/2} e_{dgj}, \tag{3.8}$$

where b_g is the effect of socioeconomic group g, $g = 1, \dots, G-1$ ($b_G = 0$, $G = 4$). The estimator of the dth small area mean can be expressed as

$$\widehat{\overline{Y}}_d^{greg3} = \widehat{\overline{Y}}_d^{wdirect} + \sum_{g=1}^{G-1} \widehat{b}_g \left(\frac{N_{dg}}{N_d} - \frac{w_{dg}}{w_d} \right) + \widehat{\beta}_1(\overline{X}_{d1} - \widehat{\overline{X}}_{d1}^{wdirect}) + \widehat{\beta}_2(\overline{X}_{d2} - \widehat{\overline{X}}_{d2}^{wdirect}),$$

where $\widehat{b}_g$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are the weighted least squares estimators obtained by (3.3).

3.4 Empirical best linear unbiased estimators

We consider estimators obtained by fitting to the sample a random effects model of the form

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + W^{-1/2}\boldsymbol{e}\,, \tag{3.9}$$

where $\boldsymbol{y} = \boldsymbol{y}_{n\times 1}$, $\boldsymbol{X} = \boldsymbol{X}_{n\times p}$, $\beta = \beta_{p\times 1}$, $\boldsymbol{Z} = \boldsymbol{Z}_{n\times D} = \operatorname{diag}(\boldsymbol{1}_{n_1}, \ldots, \boldsymbol{1}_{n_D})$ with $\boldsymbol{1}_a^t = (1, \ldots, 1)_{1\times a}$ and $\boldsymbol{W} = \boldsymbol{W}_{n\times n} = \operatorname{diag}(w_{11}, \ldots, w_{Dn_D})$ with $w_{11} > 0, \ldots, w_{Dn_D} > 0$ known. Vectors of area and household random effects, $\boldsymbol{u} = \boldsymbol{u}_{D\times 1} \sim N(0, \sigma_u^2 I_D)$ and $\boldsymbol{e} = \boldsymbol{e}_{n\times 1} \sim N(0, \sigma_e^2 I_n)$ respectively, are assumed to be independent with unknown variance components σ_u^2 and σ_e^2. Under this model, the variance-covariance matrix of $\boldsymbol{y}$ is

$$V = \sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}^t + \sigma_e^2 W^{-1}\,.$$

The individual predictions of non sampled units are

$$\widehat{Y}_{dj} = \boldsymbol{x}_{dj}\widehat{\beta} + \gamma_d^w \left(\widehat{\overline{Y}}_d^{wdirect} - \widehat{\overline{X}}_d^{wdirect} \widehat{\beta} \right),$$

where

$$\widehat{\beta} = (\boldsymbol{X}^t V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^t V^{-1} \boldsymbol{y}, \quad \gamma_d^w = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 / w_d}, \quad w_d = \sum_{j=1}^{n_d} w_{dj}, \quad d = 1, \ldots, D.$$

The *blup* of small area mean $\overline{Y}_d$ is obtained under the assumption of known variance components (see Chapter 2 of Vaillant et al (2000) for an overview of the Prediction Theory) by

$$\begin{aligned} \widehat{\overline{Y}}_d &= \frac{1}{N_d} \left(\sum_{j \in s \cap d} Y_{dj} + \sum_{j \in r \cap d} \widehat{Y}_{dj} \right) \\ &= (1 - f_d) \left[\overline{\boldsymbol{X}}_d \widehat{\beta} + \gamma_d^w \left(\widehat{\overline{Y}}_d^{wdirect} - \widehat{\overline{\boldsymbol{X}}}_d^{wdirect} \widehat{\beta} \right) \right] + f_d \left[\widehat{\overline{y}}_d + (\overline{\boldsymbol{X}}_d - \widehat{\overline{\boldsymbol{X}}}_d) \widehat{\beta} \right], \end{aligned} \tag{3.10}$$

where $\overline{\boldsymbol{X}}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \boldsymbol{x}_{dj}$, $\widehat{\overline{\boldsymbol{X}}}_d = \frac{1}{n_d} \sum_{j=1}^{n_d} \boldsymbol{x}_{dj}$ and $\widehat{\overline{y}}_d = \frac{1}{n_d} \sum_{j=1}^{n_d} y_{dj}$.

Observe that estimator $\widehat{\overline{Y}}_d$ depends on the variance components through γ_d^w and V^{-1}. By plugging in (3.10) suitable estimators of the variance components, the *empirical blup* (*eblup*) of small area mean $\overline{Y}_d$ is obtained (see Battesse et al. (1988) and Prasad and Rao (1990)). We present some estimators obtained from specific models of the type (3.9).

- *ebluph1:* It uses X_1 as auxiliary variable, so that $\boldsymbol{x}_{dj} = x_{dj1}$ and $\beta = \beta$. Estimated variance components $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$ are obtained by Henderson's method 3 (see Henderson (1953) or Searle et al (1992)). This small area estimator has been studied by Rao and Choudhry (1995).
- *eblup1:* The only difference of this estimator with respect to *ebluph1* is that β, σ_u^2 and σ_e^2 are estimated by maximizing the likelihood of model (3.9). These

maximum likelihood estimates (MLE) are calculated numerically by using the Fisher-scoring algorithm. We will check which of both variance components estimation methods (maximum likelihood or Henderson method 3) provide better small area estimator.

- *eblup2:* It uses X_1 and X_2 as auxiliary variables. Estimators of model parameters, $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\sigma}_u^2, \widehat{\sigma}_e^2$, are MLE's and they are calculated via Fisher-scoring algorithm.
- *eblup3:* It uses X_1, X_2 and G as auxiliary variables. Estimators of model parameters, $\widehat{b}_g, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\sigma}_u^2, \widehat{\sigma}_e^2$ are MLE's and they are computed by using Fisher-scoring algorithm.

4 Measures of sampling errors in simulation experiment

In order to evaluate the precision and accuracy of proposed small area estimators for estimating the average net income, $\overline{Y}_d$, K samples are drawn from the artificial population and estimations are obtained for each sample. Let $\widehat{\overline{Y}}_d(k)$ be the estimate of the mean $\overline{Y}_d$ for the small area d in the k-th replicated sample. The following standard performance criteria are considered:

1. The *average relative bias* for small area d

$$ARB_d = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\widehat{\overline{Y}}_d(k)}{\overline{Y}_d}\right)100. \tag{4.1}$$

2. The *relative mean squared error* for small area d

$$RMSE_d = \frac{100}{\overline{Y}_d}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\overline{Y}}_d(k) - \overline{Y}_d\right)^2}. \tag{4.2}$$

5 Monte Carlo simulation experiment

A C++ Builder program has been developed to extract random samples from the data file and to evaluate estimators and performance measures. The number of replications of the simulation experiment is $K = 10{,}000$. A *stratified sampling design*, with simple random sampling without replacement inside each of the strata and total sample size $n = 600$, has been used. Population sizes of strata N_h are calculated from the artificial population, and sampling weights w_h have been taken from the Spanish Family Budget Survey for a province with average size. From these two quantities, by the relation $w_h = N_h/n_h$, sample sizes inside each stratum n_h have been derived. These quantities are shown in Table 5.1.

***Table** 5.1: Sizes and weights per stratum.*

stratum	1	2	3	4	5	6
N_h	49828	52717	48051	48865	48831	51708
n_h	100	105	96	98	98	103
w_h	498.28	502.07	500.53	498.62	498.28	502.02

With the obtained sample, all estimators of the average net income $\overline{Y}_d$ are calculated for small areas defined by $D_1, \dots, D_6$. When the process of $K = 10,000$ replications is finished, efficiency measures are evaluated.

In order to clarify the role that sample size ($n = 600$) and number of small areas (D) play in the analysis of the numerical results, in Table 5.2 we present the quantities n/D for D_1–D_6.

***Table** 5.2: Average sample sizes per small areas.*

	D_1	D_2	D_3	D_4	D_5	D_6
D	8	16	32	64	128	256
$600/D$	75	37.5	18.75	9.375	4.6875	2.34375

***Table** 5.3: Means over small areas and standard deviations (in brackets) of ARB_d.*

Estimator	D_1	D_2	D_3	D_4	D_5	D_6
wdirect	99.998	100.005	100.026	100.001	100.001	100.145
	(0.009)	(0.022)	(3.167)	(0.036)	(0.107)	(0.966)
synt	100.012	100.037	100.027	100.143	100.337	100.704
	(2.012)	(2.595)	(3.127)	(4.207)	(5.606)	(7.318)
ssd	100.012	100.037	100.026	100.020	100.062	100.219
	(2.068)	(2.640)	(3.163)	(0.551)	(1.036)	(1.968)
rsynt	100.015	100.036	100.038	100.161	100.343	100.693
	(1.371)	(2.160)	(2.758)	(3.972)	(5.404)	(7.120)
greg1	100.002	99.999	100.002	100.005	100.004	100.142
	(0.008)	(0.018)	(0.025)	(0.031)	(0.107)	(0.915)
greg2	100.002	99.999	100.003	100.006	100.006	100.142
	(0.005)	(0.016)	(0.023)	(0.029)	(0.108)	(0.895)
greg3	100.002	99.999	100.002	100.004	100.001	100.145
	(0.006)	(0.017)	(0.024)	(0.030)	(0.105)	(0.966)
ebluph1	100.000	99.994	99.992	99.972	98.890	90.194
	(0.008)	(0.018)	(0.025)	(0.038)	(0.918)	(6.423)
eblup1	99.984	99.975	99.938	99.868	98.740	90.171
	(0.010)	(0.013)	(0.023)	(0.044)	(0.889)	(6.382)
eblup2	99.977	99.960	99.885	99.601	99.680	99.827
	(0.008)	(0.013)	(0.030)	(1.529)	(2.232)	(3.035)
eblup3	100.001	100.026	100.020	100.046	100.113	100.264
	(0.595)	(0.767)	(1.028)	(1.331)	(1.760)	(2.336)

In Table 5.3, means and standard deviations (in brackets) of ARB_d over small areas, that is, $\overline{ARB} = D^{-1} \sum_{d=1}^{D} ARB_d$ and $S_{ARB} = \left[D^{-1} \sum_{d=1}^{D} (ARB_d - \overline{ARB})^2\right]^{1/2}$, are listed for each variable defining small areas $D_1 - D_6$. In Table 5.4 means and standard deviations of $RMSE_d$ are given. Observe that in Table 5.3, standard deviations provide more information about the amount of bias than the mean, which is "average" bias.

***Table** 5.4: Means over small areas and standard deviations (in brackets) of $RMSE_d$.*

Estimator	D_1	D_2	D_3	D_4	D_5	D_6
wdirect	1.320	1.873	2.697	3.805	5.420	6.687
	(0.083)	(0.152)	(1.647)	(0.488)	(0.917)	(1.000)
synt	1.745	2.101	2.659	3.324	4.572	6.065
	(0.875)	(1.486)	(1.631)	(2.582)	(3.263)	(4.159)
ssd	1.801	2.148	2.693	3.267	4.319	5.575
	(0.884)	(1.497)	(1.646)	(0.386)	(0.726)	(1.275)
rsynt	1.187	1.707	2.303	3.172	4.465	5.823
	(0.655)	(1.306)	(1.508)	(2.392)	(3.061)	(4.157)
greg1	1.107	1.558	2.211	3.074	4.246	5.087
	(0.089)	(0.160)	(0.295)	(0.502)	(0.927)	(0.925)
greg2	1.068	1.495	2.118	2.931	4.008	4.756
	(0.091)	(0.162)	(0.314)	(0.524)	(0.938)	(0.890)
greg3	1.033	1.446	2.048	2.827	5.132	6.687
	(0.092)	(0.165)	(0.320)	(0.535)	(0.900)	(1.000)
ebluph1	1.107	1.558	2.211	3.218	9.868	28.056
	(0.008)	(0.018)	(0.025)	(0.038)	(0.918)	(6.423)
eblup1	1.110	1.561	2.205	3.214	9.860	27.697
	(0.092)	(0.158)	(0.287)	(0.575)	(3.362)	(8.234)
eblup2	1.067	1.503	2.132	2.537	3.406	4.448
	(0.093)	(0.162)	(0.303)	(0.722)	(0.998)	(1.421)
eblup3	0.913	1.240	1.642	2.220	2.976	3.842
	(0.241)	(0.314)	(0.380)	(0.582)	(0.851)	(1.263)

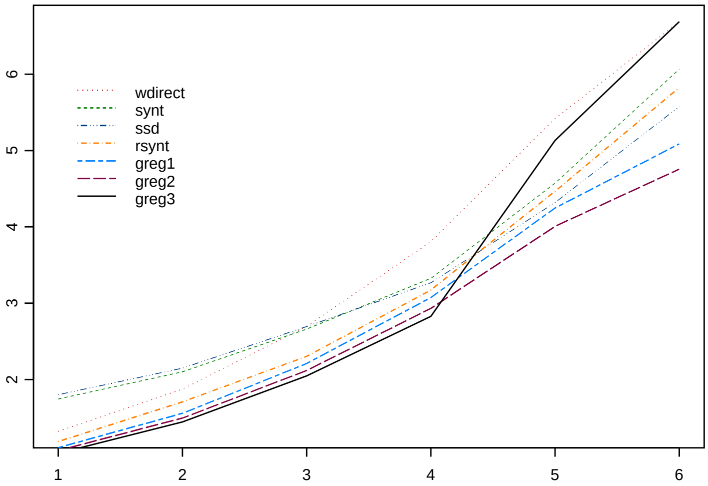

***Figure 1**: $\overline{RMSE}$ of design-based and greg estimators, for D_1-D_6.*

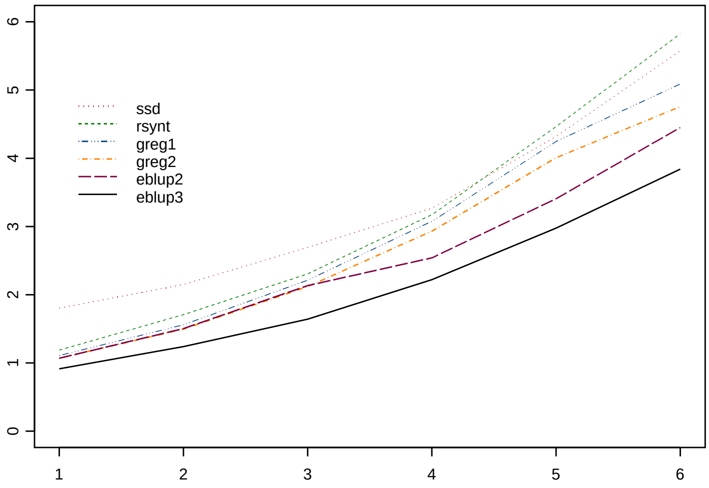

Figure 2: *$\overline{RMSE}$ of best estimators, for D_1-D_6.*

The most interesting estimators are the ones with better performance for $D_4 - D_6$, whose average area sample sizes are smaller than 10 units. We can observe better the evolution of $\overline{RMSE}$ for the small areas defined by D_1-D_6 in Figures 5.1 and 5.2. Figure 5.1 shows the $\overline{RMSE}$ for design-based and *greg* estimators, while Figure 5.2 represents the same efficiency measure only for the group of estimators with a reasonable behaviour. We make a comparison of estimators one by one.

The w-direct estimator (*wdirect*) uses just data of the target variable in the given small area. As expected, it behaves well with respect to bias in all cases; see that the standard deviation remains small even for D_6. However, its behaviour is not very good in relative mean squared error, since it increases monotonically, differing considerably from the group of good estimators from D_2 on. Thus, the use of *wdirect* can only be recommended for sample sizes larger than 40.

The basic synthetic estimator (*synt*), which uses as auxiliary information just the socioeconomic group G, is relatively stable in $\overline{RMSE}$ with respect to n/D. But we can see in Table 5.3 that it is quite biased in all cases, being the standard deviation of the ARB_d unacceptably large for D_4–D_6.

The sample size dependent estimator (*ssd*) is a composition of the direct and the basic synthetic estimator. Surprisingly, this estimator gets quite good results in both performance measures, relative mean squared error and average relative bias, being its $\overline{RMSE}$ comparably to the *greg1* and *greg2* estimators, and being its bias smaller that all *eblup* estimators, as expected, and also better than both synthetic and generalized regression synthetic estimator. This estimator is a good alternative, either when no more than a grouping variable is available as auxiliary information, or when it is not found a good model fitting the data.

The generalized regression synthetic estimator (*rsynt*) clearly improves the numerical results of *wdirect* and *synt* estimators, improving the direct estimator even for larger areas. The reason is that model (3.5) fits reasonably well to data. However, it has comparable amount of bias to the basic synthetic estimator (see the standard deviation).

Comparing generalized regression estimators, we can see in *greg1* and *greg2* how extra information provide a decrease in $\overline{ARB}$ and $\overline{RMSE}$, except when we arrive to *greg3*, whose $\overline{RMSE}$ increases quickly when going from D_4 on. We can look at case D_6 to explain this phenomenon. For small areas defined by D_6, model (3.8) from *greg3* has $D + (G - 1) + 2 = 261$ regression parameters. Since sample size is 600, there are $600/261 \approx 2.3$ observations per parameter. This ratio is too small in order to estimate model parameters with low standard deviations. Under these conditions, these model is not stable and therefore its use is not recommended.

Now we compare model-based estimators. Looking at Table 5.4, we see that *ebluph1* and *eblup1* increase considerably their $\overline{RMSE}$ for D_5 and D_6, being *greg1* (based in the same model but with fixed effects) much preferable. Since sampling fractions n_d/N_d are close to zero, the eblup estimators are approximately given by

$$\widehat{\overline{Y}}_d \cong \overline{\boldsymbol{X}}_d\widehat{\beta} + \gamma_d^w\left(\widehat{\overline{Y}}_d^{wdirect} - \widehat{\overline{\boldsymbol{X}}}_d^{wdirect}\widehat{\beta}\right). \tag{5.3}$$

The problem here is that the variability of X_1 in the small number of observations is not sufficient to estimate with precision the variability between households σ_e^2 and between areas σ_u^2. This fact causes a strong negative bias on the synthetic part of the *eblup1* estimator, $\overline{\boldsymbol{X}}_d\widehat{\beta}$. In fact, for D_6, its $\overline{ARB}$ is 85.897, and its standard deviation is 1.587. This synthetic estimator is corrected by the second term on the right of (5.3), but this term is affected by the same problem, making the correction rather poor. In this experiment, estimators based on linear models with small area random effects appear to be very sensible to goodness–of–fit, providing worse results than estimators based on linear models with small area fixed effects, but also worse than design-based estimators which do not make use of any covariate. It is interesting to note that Fisher–scoring algorithm (*eblup1*) is slightly better that Henderson's method 3 (*ebluph1*).

Estimators *greg2* and *eblup2* rely on the same linear regression models. The difference is that small area effects are fixed in the first case and random in the second case. We can see that for sample sizes smaller than 20, *eblup2* has less $\overline{RMSE}$. This indicates that as soon as selected model fits better to data, their corresponding estimators perform more efficiently.

The same occurs with *greg3* and *eblup3*. Since the model of *eblup3* is very close to the real one, this estimator presents the best numerical results in the simulation experiment. Mean squared errors remain small even in the case D_6.

6 Summary and Conclusions

In practical applications of the estimation of means and totals for small areas, statisticians needs recommendations about what type of estimator to use, when it is better a model-based approach than a model-assisted one, or how estimators are affected by the ratio *sample size/number of domains*. Theoretical properties of estimators give answers to these questions under ideal conditions, i.e. if sufficient hypotheses are fulfilled and/or sample sizes are large enough. However, in practice, models do not perfectly fit to data and sample sizes are small. Then simulation studies play an important role to gain intuition and obtain conclusions about the behaviour of estimators. The simulation study presented in this paper has tried to reproduce artificially populations and sampling designs appearing in official statistics, so that our conclusions are valuable for applied statisticians and can be taken into account by Statistical Offices.

From the developed simulation study, under the artificial population generated as described in Section 2, we can extract the following conclusions:

1. If there is an informative grouping variable available, a good choice for estimating means of areas with average sample sizes smaller than 20 is the *ssd* estimator.
2. If there is available at least one "good" covariable, its use is always recommended.
3. Best numerical results are obtained for those estimators with models fitting "well" to data. A bad model produces a bad estimator.
4. When a good model is not found, it is better to use models with fixed effects. Among estimators based in models with random effects, the worst numerical results are obtained by *ebluph1* and *eblup1*. These two estimators behave acceptably well only for average sample sizes greater than 10.
5. If a good model is available, then random effects are clearly preferred to fixed effects for sample sizes smaller than 20. Note that there is a significant decrease of relative mean squared error when passing from *greg2* to *eblup2* or from *greg3* to *eblup3*.

Acknowledgements

This work has been done in the set up of a R&D contract with Instituto Nacional de Estadística (INE). It has been also partially supported by grant BMF2003-04820 of the Spanish Ministery of Science and Technology. Ideas appearing in the manuscript come from the collaboration of authors with INE and with EURAREA team. EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs, IST-2000-5.1.8, 2001-2003) is a research project funded by EUROSTAT under the Fifth Frame work Programme.

References

Battesse, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Drew, J. D., Singh, M. P. and Choudhry, G. H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 17-47.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.

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.

Rao, J. N. K. and Choudhry, G. H. (1995). Small area estimation: overview and empirical study. *Business Survey Methods* (Cox, Binder, Chinnappa, Christianson, Colledge, Kott, eds.). John Wiley, 527-542.

Särndal, C., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Searle, S. R., Casella, G. and McCullogh, C. E. (1992). *Variance Components*. John Wiley. New York.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference. A Prediction Approach*. John Wiley. New York.