STATISTICS IN TRANSITION-new series, August 2011
Vol. 12, No. 1, pp. 57—80

ESTIMATION OF AVERAGE INCOME IN CUBAN MUNICIPALITIES

Nestor Arcia Montes De Oca[1]

ABSTRACT

This article asserts that diversification to produce small area statistics for different fields in Cuban society should be a priority for the National Statistics Office. The key question about small area estimation is how to obtain reliable local statistics when the sample data contain too few observations for statistical inference of adequate precision. The social research presented here is focused on finding small area estimates which are more precise than the direct estimates of monthly mean income for people aged 15 and over at a municipal level. In this case, all 169 Cuban municipalities are considered small areas of interest.

The empirical results obtained from this application are only intended to provide a first impression of the usefulness of applying small area estimation methods in Cuba. This study yields more precise estimates than the direct estimates for small areas/domains, even though in Cuba, as in any other developing country, the search for suitable auxiliary variables is used to "borrow strength" from neighbouring areas or domains may frequently be an important limitation.

Key words: small area estimation, borrow strength.

1. Introduction

In recent years, the academic world has taken an increasing interest in the analysis of regional economic disparities which represent a serious challenge to achieving national economic growth and, thus, social cohesion. In Cuba, this analysis was particularly apposite after the collapse of the Soviet Union in the late 1980s and early 1990s, when the Cuban economy experienced a severe crisis. As part of its overall response to the crisis, the Cuban government initiated a series of monetary and market reforms which were designed to provide material incentives for economic activities. These included: expanding international tourism and joint

[1] National Statistics Office. Email: nestor@one.cu.

ventures; legalising the possession of US dollars - making it possible for Cubans to receive money from relatives or friends living abroad; permitting self-employment, agricultural markets and other markets (e.g. handicrafts) and others. While these reforms contributed to the subsequent economic recovery, they also resulted in an increase in economic inequality, even though salary scales in Cuba still remain fairly egalitarian. This fact has increased interest in producing regional statistical information and has stimulated research on income distribution, poverty and social exclusion at the small area level in Cuba.

In this respect, this article proposes a new small-area application intended to produce small area estimates which are more precise than direct estimates of *monthly mean income for those aged 15 and over at a municipal level*. The 169 Cuban municipalities are considered as the small areas of interest. These direct estimates are obtained from the second national survey on health risk factors and non-communicable chronic ailments (HRF) which was conducted during November-December 2000 by the National Institute of Hygiene, Epidemiology and Microbiology and the National Statistics Office (ONE) which, in turn, was represented by the Centre for Population and Development Studies (CEPDE, Spanish acronym) and the Social Statistics Department. In this survey, out of the 169 Cuban municipalities, 158 are represented in the sample and the remaining 11 are not sampled at all. Likewise, the direct estimates for some of these 158 sampled municipalities cannot be produced because they have unacceptably large variances. The challenge is to find plausible strategies for improving the direct estimates for the 158 sampled municipalities, when needed, and also to find acceptable estimates for the 11 non-sampled municipalities. Such a strategy may consist of using the same national sample size as the HRF survey and producing the municipal estimates by small area estimation (SAE) methods. If this strategy proves easy to apply, policy makers could use it to acquire more accurate information at a municipal level and this could be considered as an initial finding to apply to other indicators in future surveys.

In addition to the direct estimator, two small area model-based estimators are proposed in this article: (*a*) the synthetic-regression estimator; and (*b*) the empirical best linear unbiased prediction (EBLUP) estimator. The basic unit level model (Rao, 2003) is used to derive each of these two small area estimators. Using HRF survey data, it illustrates how the small area estimates derived from such a model could potentially be useful to improve the efficiency of the direct small area estimates. The MSE estimator of each estimator is used as a measure of uncertainty. Since the real population data are not available, two simulation experiments are used - a model-based simulation and a design-based simulation - to compare the performance of these three estimation methods. The performances of the MSE estimators for the best small area estimator are also compared to each other.

This article is structured as follows. Section 2 describes the sampling plan and the method of estimation. Section 3 describes the SAE methods (synthetic-

regression estimator and EBLUP) derived from the basic unit level model, which will be compared with the direct estimator. The results obtained from the basic unit level model are presented in Section 4. The comparison between the three estimators is based on the results obtained from model and design-based Monte Carlo simulation studies (Section 5). Section 6 presents the results obtained from the application using HRF data. Section 7 gives some final remarks and suggestions for further research.

2. Sampling design

This section describes the sampling design of the HRF survey, as well as the direct estimator of monthly mean income and its MSE estimator. The performance of this estimator will be compared to the two SAE estimators described later in this article (Section 3).

The HRF sample was drawn from the Cuban sampling frame for population surveys (ONE, 2004). A three-stage cluster probability sample was selected from urban areas. Fourteen provinces and the special Isle of Youth municipality were used as strata. Primary sampling units (PSU) were the sampled geographical areas (AGEM with an average 180 housing units (ONE, 2004), drawn with probability proportional to size, using the total number of houses with permanent residents within each stratum as size measure. The second sampling units (SSU) were blocks, which have at least 30 housing units and they were also selected with probability proportional to size using the total number of houses with permanent residents within each AGEM as size measure. Tertiary and last sampling units (TSU) were the sections (5 housing units on average), which were drawn with equal probability. A block is drawn within each selected AGEM, and finally two sections were drawn in each selected block, giving 10 houses in each AGEM. All adults aged 15 or over were interviewed in each selected house. The sampling design assured equal probability within each province, which allowed the acquisition of accurate and reliable information of the main health indicators at provincial and national level. Assuming a 95% confidence interval and 5% survey non-response rate, it is guaranteed a proper sample size to obtain the core outputs from the survey at provincial and national level. The province allocation was made with probability proportional to size that is given by the population of people aged 15 or older in each province. This sampling design leads to a sample size distribution for the 158 sampled municipalities that ranges between 16 and 1148 individuals, with no individuals for the 11 non-sampled municipalities (Figure 1A).

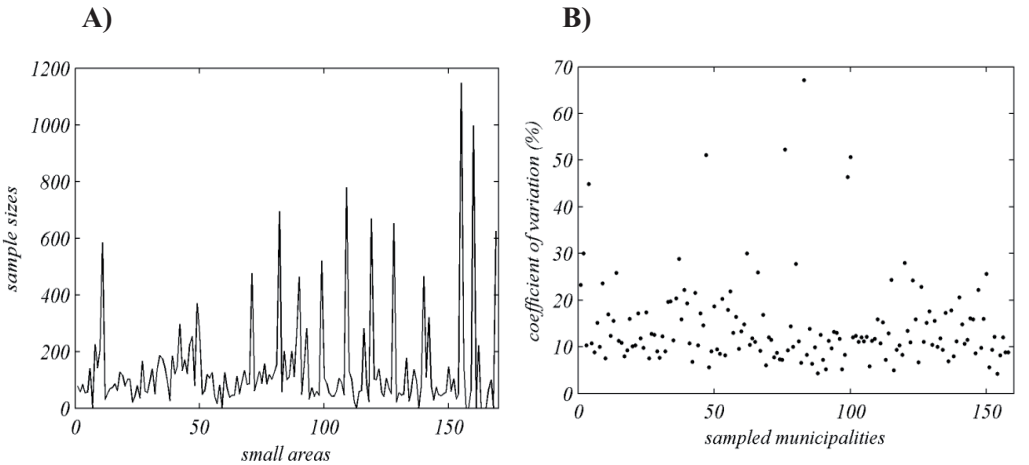

Figure 1. A) Sample size distribution between the small areas Cuban municipalities. B) Sample size vs. the coefficient of variation of the direct estimates for each sampled municipality.

Figure 1B shows that the direct estimates are not acceptable for some municipalities, particularly for those with a coefficient of variation greater than 20% (Sarndal et al., 1992). This confirms the need to find methods of producing more precise estimates than the direct estimates for such municipalities. The sampling plan yields a self-weighting sample within the stratum (province) which means that every individual has the same probability of being included in the sample within each province. Thus, all the municipalities within a specific province will also have the same individual sample weight. Taking into account the self-weighting sample property, the direct estimator of the monthly mean income in the municipality *i* is given by:

$$\hat{\bar{y}}_i = \frac{\sum_{j,k,l,h\in S_i} w_t y_{jklh}}{\sum_{j,k,l\in S} w_t} = \frac{\sum_{j,k,l,h\in S_i} y_{jklh}}{n_i} \tag{1}$$

where n_i is the sample size of the municipality *i* and $j,k,l,h \in S_i$ denotes the sampled individual *h* within the section *l* within the block *k* within the AGEM *j* within the municipality *i*. By using this sampling plan, the direct estimator in each municipality coincides exactly with the sample mean. This estimator will be compared with the two small area estimators that are used later in this article.

The calculation of the variance estimate of the direct means under the HRF sampling design is difficult since it requires the knowledge of the first and second-order inclusion probabilities. For simplicity, the following variance estimator is used

$$\hat{MSE}\left(\hat{\bar{y}}_i\right) = \frac{n_t\left(N_t - n_t\right)\left(N_i - 1\right)\left(n_i - 1\right)}{{n_i}^2 N_t\left(N_t - 1\right)b_i} s_{yi}^2 \tag{2}$$

where N_t and n_t are the population and sample size in the province t respectively, n_i and N_i are the sample and population size of the municipality i respectively, $b_i = \frac{n_t}{N_t}\left(N_i - 1\right)\left[1 - \frac{\left(N_t - n_t\right)}{\left(N_t - 1\right)n_i}\right]$, and $s_{yi}^2 = \frac{1}{\left(n_i - 1\right)} \sum_{j,k,l,h \in S_i} \left(y_{jklh} - \hat{\bar{y}}_i\right)^2$.

The estimator (2) coincides with the MSE estimator of the direct mean for each municipality i when a simple random sampling without replacement (SRSWOR) is considered in the province t. The implication of ignoring clustering may lead to an underestimation of the true MSE and consequently the confidence intervals will be narrower than expected. In this application we accepted to take such a risk because the MSE estimates of the direct means obtained under the HRF sampling design will always be greater than the MSE estimates of the direct means under a SRSWOR plan, i.e. the design effects will always be greater than 1. Therefore, if the precision of a small area estimate is found to be better than the precision of the direct mean measured under SRSWOR plan, it will also be better than the precision of the direct mean measured under the HRF sampling design.

3. Small area estimators

This section aims to describe the SAE methods and their MSE. The direct estimator of monthly mean income in each of the 158 sampled municipalities (1) will be compared with the two small area estimators derived from the basic unit level model (synthetic-regression and EBLUP estimators).

The direct estimator and its MSE were both described in Section 2. By avoiding the notation system employed for each sampling unit, the direct estimator (sample mean) can be expressed as: $\hat{\bar{Y}}_{iD} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$, where the subscript i and j identify the municipality and the individual, respectively. The MSE estimator ($mse\left(\hat{\bar{Y}}_{iD}\right)$) was given in (2). The general synthetic estimator can be

expressed as: $\hat{\bar{Y}}_{iSR} = \bar{X}_i^t \hat{\beta}$ (3), where $\bar{X}_i$ is the *ith* population proportion of x_{ij}'s, and $\hat{\beta}$ may be estimated by using any likelihood estimation methods or a moment method. The synthetic estimator (3) can be considered a model-based estimator and the MSE of $\bar{X}_i^t \hat{\beta}$ can then be estimated by adopting this approach. Using the assumptions of the basic unit level model, the unknown true mean $\bar{Y}_i$ can be considered itself as a random variable with variance σ_υ^2 and expectation $\bar{X}_i^t \hat{\beta}$. The synthetic estimator $\bar{X}_i^t \hat{\beta}$ can also be treated as an independent random variable with the same expectation. Then, the MSE of $\bar{X}_i^t \hat{\beta}$, under such assumptions, is given by Goldring et al. (2005) as:

$$MSE\left(\bar{X}_i^t \hat{\beta}\right) = E\left(\bar{X}_i^t \hat{\beta} - \bar{Y}_i\right)^2 = \bar{X}_i^t V\left(\hat{\beta}\right) \bar{X}_i + \sigma_\upsilon^2 \tag{4}$$

An estimate of the MSE of $\bar{X}_i^t \beta$ can be obtained by substituting the restricted iterative generalized least squares (RIGLS) estimates $\hat{\sigma}_\upsilon^2$ and $\hat{V}\left(\hat{\beta}\right)$ for σ_υ^2 and $V\left(\hat{\beta}\right)$ respectively (4) (Goldstein, 1989). The EBLUP estimator is a weighted average of the "survey regression" estimator $\left[\hat{\bar{Y}}_{iD} + \left(\bar{X}_i - \bar{x}_i\right)^t \hat{\beta}\right]$ and the synthetic estimator $\bar{X}_i^t \hat{\beta}$. It is given by (Rao, 2003):

$$\hat{\bar{Y}}_i^{EBLUP} = \hat{\gamma}_i \left[\hat{\bar{Y}}_{iD} + \left(\bar{X}_i - \bar{x}_i\right)^t \hat{\beta}\right] + \left(1 - \hat{\gamma}_i\right) \bar{X}_i^t \hat{\beta} \tag{5}$$

where $\hat{\gamma}_i = \dfrac{\hat{\sigma}_\upsilon^2}{\hat{\sigma}_\upsilon^2 + \dfrac{\hat{\sigma}_e^2}{n_i}}$, $\left(\hat{\bar{Y}}_{iD}, \bar{x}_i\right)$ are the *i-th* area sample means, and $\hat{\sigma}_\upsilon^2$, and $\hat{\sigma}_e^2$ are the variances at level-2 and level-1 respectively. They are estimated using the RIGLS method. A set of MSE estimators can be used for the EBLUP estimator (5). Two such MSE estimators, the MSE estimator given by Prasad and Rao

(1990) (PR EBLUP MSE) and the MSE estimator proposed by Jiang, Lahiri and Wan (2002) (JLW EBLUP MSE), will be used in this article. The leading term in both MSE estimators is $g_{1i}\left(\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)=\hat{\gamma}_i\left(\frac{\hat{\sigma}_e^2}{n_i}\right)$ (Prasad and Rao, 1990; Jiang, Lahiri and Wan, 2002).

4. Basic unit level model

Both the synthetic estimator and EBLUP estimator are derived from the basic unit level model. The basic unit level model was selected due to the continuous nature of the response variable (total individual monthly income) and because the auxiliary information is available at a small area level, which is an important prerequisite to produce small area estimates derived from this model. Of the many possible covariates available in the HRF questionnaire, those for which area means were available from the 2002 census results were chosen: sex, age, marital status and educational level. These four covariates are included into the model by applying a backward elimination procedure. The RIGLS method is used to estimate the fixed and random parameters of the selected model. Table 1 gives the summary statistics for three multilevel models of income behaviour in Cuba. The four level-1 predictors - sex, marital status, education and age - are specified as a set of six dummy indicator variables. The reference person is a single woman aged between 15 and 34 years old who lives in Havana City and has a low educational level (less than a high school education).

Table 1. Multilevel estimates for models of income behaviour

	Estimates
Fixed effects	
Level 1	
Intercept	122.760(14.808)
Female	-
Male	82.279(6.977)
Single	-
Married	22.943(7.381)
Less than high school	-
High school or above	125.383(7.532)
15-34	-
35-54	75.411(8.225)
55-74	48.322(10.078)
75+	29.086(15.663)
Level 2	
Havana City	-
Pinar del Rio	-73.122(19.735)
Havana Province	-96.729(19.127)
Matanzas	-101.810(20.413)
Villa Clara	-97.031(20.304)
Cienfuegos	-76.086(21.453)
Sancti Spiritus	-82.827(20.915)
Ciego de Avila	-98.394(21.365)
Camaguey	-87.239(21.291)
Las Tunas	-123.486(21.758)
Holguin	-99.007(20.712)
Granma	-136.161(20.837)
Santiago de Cuba	-106.606(22.646)
Guantanamo	-126.880(23.836)
Isla de la Juventud	-84.514(34.142)
Random effects	
Level 1	
Intercept, σ_e^2	276050.400(2589.162)
Level 2	
Intercept, σ_υ^2	562.424(268.720)

Standard deviation for each regression estimates are shown in brackets.

To test normality, a normal Q-Q plot and the Shapiro-Wilk's test for standardised residuals of the fitted model were examined. From the figure below, neither standardised municipal nor individual residuals are close to the $y = x$

line, which suggests that the normality assumption for the residuals does not hold. The Shapiro-Wilk's test also rejected the null hypothesis of normality for both residuals (p-value=0.0000).

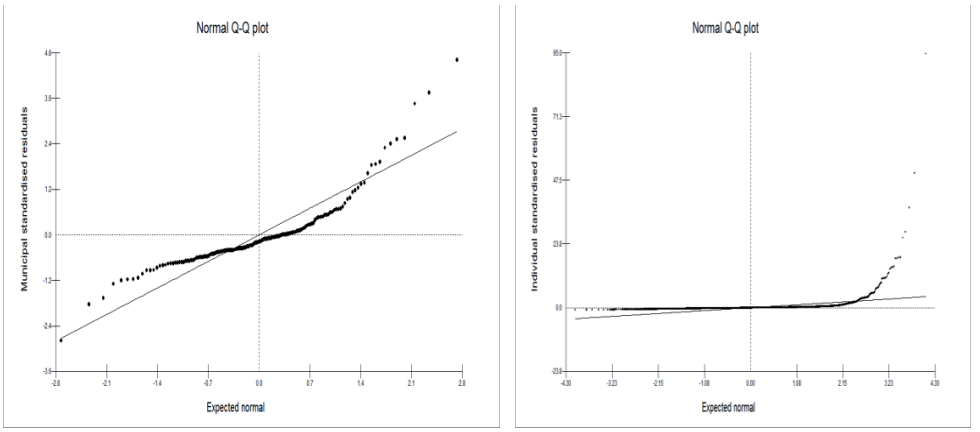

Figure 2. Municipal standardised residuals vs. expected normal (left-side) and individual standardised residuals vs. expected normal (right-side).

We relax the usual normality assumption for residuals in order to examine the robustness of the small area estimates derived from such a model. This means that the effects of the residuals' non-normality may have an impact on the small area estimators and such an impact may be even worse for the MSE small area estimators. This will be examined later in the simulation study (Section 5). Finally, the logarithmic transformation is widely used in models for income because the logarithms of values generated from a positive asymmetric distribution are generally more "normal" than the raw values. However, it was decided not to use such a transformation because almost 30% of individuals in the data set have reported a total income equal to zero, which might introduce a bias into the estimation of the small area estimators. Note that the logarithmic transformation effectively controls the influence of raw-scale outliers, but is then susceptible to log-scale outliers (e.g. either values near zero or variables that contain a significant proportions of zeros).

5. Simulation studies

This section details the results from two simulation experiments that enable the comparison of the three SAE methods described in Section 3. These simulation experiments also allow the performance of the MSE estimators for the best small area estimator to be evaluated. The first is a model-based simulation in which a small area population (municipality population) and sample data are generated from a distribution model. In this case, small area population and

sample data are based on the basic unit level model, assuming normal distributions for the two random effects. The second is a design-based simulation in which a fixed population is generated from existing samples and sample data are drawn from this fixed population. In this case, the fixed population is generated using the HRF sample data. The way to generate the population for the model-based and design-based simulation experiments are outlined in Section 5.1 and 5.2 respectively. The multilevel for Window (MLwiN) statistical software is used to implement all the small area estimators and their MSE estimators (see Appendice 1).

Since the residuals do not show evidence of normality (see the Normal Q-Q plot in Section 4), if there are different estimators' patterns for these two different simulation experiments, those found in the design-based simulation will be assumed to be more accurate and realistic than the other one. Therefore, the results from the model-based simulation will only be used for a simple empirical purpose. These, along with those from the design-based simulation, will enable an appraisal of the extent to which the small area estimators and their MSE estimators change under different population assumptions.

The Average Absolute Relative Bias (*AARB*) and the Average Mean Root Square Error (*ARMSE*) are used to evaluate the estimators' performance in the two simulation processes:

$$AARB = m^{-1}\sum_{i=1}^{m}\left|R^{-1}\sum_{r=1}^{R}\left(\frac{\hat{\bar{Y}}_{i(r)}}{\bar{Y}_{i(r)}}-1\right)\right| \qquad ARMSE = m^{-1}\sum_{i=1}^{m}\sqrt{R^{-1}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_{i(r)}-\bar{Y}_{i(r)}\right)^{2}}$$

where $\hat{\bar{Y}}_{i(r)}$ and $\bar{Y}_{i(r)}$ are the small area estimator and the true mean for the municipality *i* at the simulation *r*, respectively. The gain in efficiency connected to each small area estimator is evaluated using the ratio of its *ARMSE* to the *ARMSE* of certain estimator that is used as benchmark. In particular, all estimators are compared with the direct estimator, and this ratio is denoted as $AEFF_{dir}$. Empirical results on model-based and design-based simulation processes are given in Sections 5.1 and 5.2 respectively.

5.1 Model-based simulation

This section compares the performance of the three small area estimators (direct, synthetic, and EBLUP) described earlier in Section 3. This comparison is made through a model-based simulation, by using the two evaluation measures (*AARB* and *ARMSE*) mentioned in Section 5. The population values $y_{ij(r)}$ were independently generated for each of the 158 municipalities at the simulation *r* from the following basic unit level model:

$$y_{ij(r)} = 122.76 + 82.279x_{ij1(r)} + 22.943x_{ij2(r)} + 125.383x_{ij3(r)} + 75.411x_{ij4(r)} + 48.322x_{ij5(r)}$$
$$+ 29.086x_{ij6(r)} - 73.122x_{i7} - 96.729x_{i8} - 101.81x_{i9} - 97.031x_{i10} - 76.086x_{i11} - 82.827x_{i12}$$
$$- 98.394x_{i13} - 87.239x_{i14} - 123.486x_{i15} - 99.007x_{i16} - 136.161x_{i17} - 106.606x_{i18}$$
$$- 126.880x_{i19} - 84.514x_{i20} + \upsilon_i + e_{ij}$$

where the area effects υ_i and individual effects e_{ij} were independently drawn from $N(0,\sigma_\upsilon^2)$ and $N(0,\sigma_e^2)$ respectively, the values $\sigma_\upsilon^2 = 562.424$ and $\sigma_e^2 = 276050.4$ were obtained from the model fitted to the HRF data (Table 1). For each simulation *r*, the level-1 variables $\left(x_{ij1(r)}, x_{ij2(r)}, x_{ij3(r)}, x_{ij4(r)}, x_{ij5(r)} \text{ and } x_{ij6(r)}\right)$ are generated from Bernoulli distribution with probability equal to their corresponding population proportions and using a population size of the municipality *i* equal to N_i. This population size is obtained from the individual's sample weight, $N_i = n_i \frac{N_t}{n_t}$, with n_i the sample size of the municipality *i*, and N_t and n_t are the population and sample size in the province *t* (the province where the individual lives), respectively. The level-2 variables $\left(x_{i7}, ..., x_{i20}\right)$ indicate which province each individual belongs to. The population mean for each municipality at the simulation *r* was then calculated as $\bar{Y}_{i(r)} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij(r)}$. This manner of calculating the population size for each small area guarantees that the population size of the pseudo-population coincides with the real population size of the population of interest (Cuban population aged 15 and over). The simulation itself was model-based, so the population values $y_{ij(r)}$ were independently regenerated at each simulation and an independent sample for each municipality was then generated from these population values each time. Within each municipality, a SRSWOR design was used to generate the sample with the sample size n_i equal to the sample size obtained from the HRF survey. The same procedure was repeated 200 times ($r = 1,...,200$). For each simulation, the three estimators (direct, synthetic, and EBLUP) were computed, along with their MSE estimators. Over the 200 simulations, the *AARB* and *ARMSE* were calculated as in Section 5. Table 2 contains the values of *AARB*, *ARMSE*, and $AEFF_{dir}$ obtained for the direct estimator, the synthetic estimator, and the EBLUP estimator:

Table 2. Performance indicators of the three small area estimates.

Estimators	*AARB*	*ARMSE*	$AEFF_{dir}$ (%)
Direct	0.019	57.90	100
Synthetic	0.018	27.08	46.77
EBLUP	0.015	25.56	44.14

Table 2 shows that the synthetic and EBLUP estimates perform significantly better than the direct estimates, leading to less than 100% $AEFF_{dir}$ values. This gain in efficiency of about 55% ($\approx 100 - 44.14$) may be due to the use of different auxiliary information from the Cuban census. In terms of bias, all three small area estimators perform very well and there is no significant difference between them. The performance of each small area estimate can be also seen graphically for each of the 158 sampled municipalities (Figure 3A). The sampled municipalities are arranged in ascending order, according to the true mean of the monthly mean income $\overline{Y}_i = \frac{1}{200}\sum_{r=1}^{200} \overline{Y}_{i(r)}$. The true mean for each municipality is also plotted. In general, there is little difference between the three small area estimates.

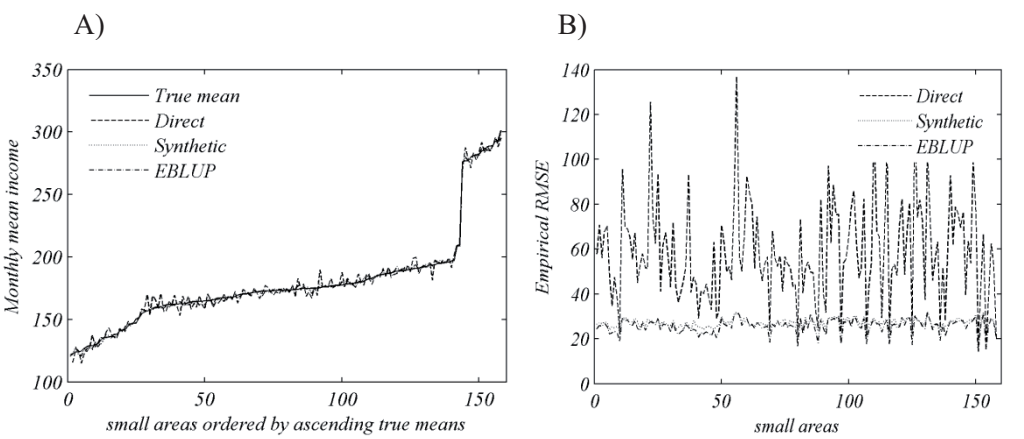

Figure 3. A) Estimated means of each small area estimates and the true mean for each of the 158 sampled small areas. B) Empirical RMSE of each small area estimates for each of the 158 sampled small areas.

Figure 3B represents the performance of all small area estimators in terms of its accuracy. The *RMSE* of the EBLUP estimate for each small area is smaller than the direct *RMSE* and slightly better than the synthetic *RMSE*. Therefore, the EBLUP estimator is more efficient than the direct estimator and is slightly more efficient than the synthetic estimator.

5.1.1 Assesing the perfromance of the eblup mse estimators

Since the EBLUP estimator emerged as the best performer in the previous section, this section aims to assess the performance of the EBLUP MSE estimators. It will compare the performance of the two EBLUP MSE estimators described in Section 3, the PR EBLUP MSE estimator and the JLW EBLUP MSE estimator. The following two measures are used to evaluate the performance of the EBLUP MSE estimators:

$$AARB = m^{-1}\sum_{i=1}^{m}\left| R^{-1}\sum_{r=1}^{R}\left(\frac{mse_{*}\left(\hat{\bar{Y}}_{i(r)}^{EBLUP}\right)}{mse_{EMP}\left(\hat{\bar{Y}}_{i}^{EBLUP}\right)} - 1 \right) \right|$$

$$ARMSE = m^{-1}\sum_{i=1}^{m}\sqrt{R^{-1}\sum_{r=1}^{R}\left(mse_{*}\left(\hat{\bar{Y}}_{i(r)}^{EBLUP}\right) - mse_{EMP}\left(\hat{\bar{Y}}_{i}^{EBLUP}\right)\right)^{2}}$$

where $mse_{EMP}\left(\hat{\bar{Y}}_{i}^{EBLUP}\right) = R^{-1}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_{i(r)}^{EBLUP} - \bar{Y}_{i(r)}\right)^{2}$ is the empirical MSE and will be used as benchmark to compare the performance of each EBLUP MSE estimator. The symbol * refers to the type of EBLUP MSE estimator, i.e., PR, and JLW. As in the case of the small area estimator, the gain in efficiency connected to EBLUP MSE estimator is also evaluated, using the ratio between the corresponding *ARMSE*, that is $AEFF_{JLW} = \frac{ARMSE_{PR}}{ARMSE_{JLW}}$. The Average Relative Bias (*ARB*) is also used to provide a better understanding of whether the given EBLUP MSE estimators tend to underestimate the MSE or not,

$$ARB = m^{-1}\sum_{i=1}^{m} R^{-1}\sum_{r=1}^{R}\left(\frac{mse_{*}\left(\hat{\bar{Y}}_{i(r)}^{EBLUP}\right)}{mse_{EMP}\left(\hat{\bar{Y}}_{i}^{EBLUP}\right)} - 1 \right)$$

Table 3 shows the values of *AARB*, *ARB*, and $AEFF_{JLW}$ for each of the EBLUP MSE estimators. In terms of bias, both the mse_{PR} and the mse_{JLW} have a good performance. The mse_{PR} slightly underestimates the true MSE (4.1%), whereas the mse_{JLW} overestimates the true MSE in only a 4.8%. This similar behaviour for both EBLUP MSE estimators is also corroborated for each small area (Figure 4). It can be seen that the values of both EBLUP MSE estimators are concentrated on the straight line.

Table 3. Performance of each EBLUP MSE estimators.

Estimators	*AARB*	*ARB*	$AEFF_{JLW}$ (%)
mse_{JLW}	0.048	0.01	100
mse_{PR}	0.041	-0.02	90.80

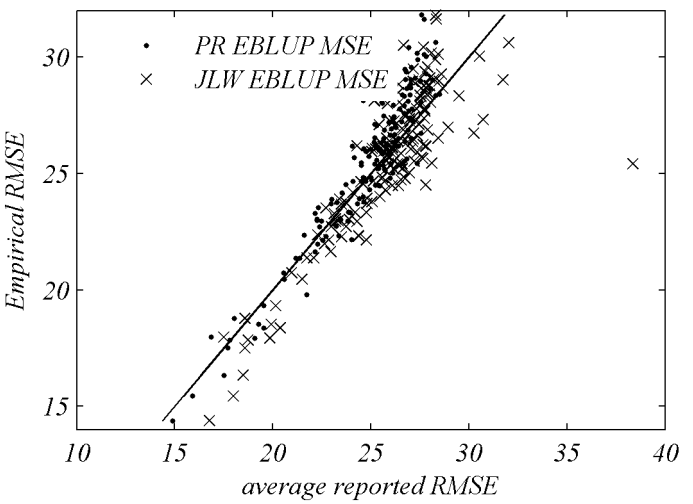

Figure 4. Empirical *RMSE* of the EBLUP estimator versus averaged reported *RMSE* of the EBLUP MSE estimators (mse_{PR} and mse_{JLW}) for each small area.

In terms of accuracy ($AEFF_{JLW}$), the mse_{PR} performs slightly better than the mse_{JLW}, offering a gain in efficiency of about 10% ($\approx 100 - 90.80$). For this reason, the mse_{PR} emerges as the most appropriate of the two EBLUP MSE estimators. However, this conclusion must be taken with caution because we assume normality for both residuals to generate the population, even though the residuals obtained from the model fitted to the HRF data did not show evidence of normality (see the Normal Q-Q plot in Section 4).

5.2 Design-based simulation

This section aims to compare the performance of three small area estimators through a design-based simulation and will also assess the performance of the MSE estimators for the best small area estimator. Due to the fact that individual income is not measured by the Cuban census, a pseudo-population was generated

using the HRF data and then samples were drawn using the original sampling design. The population values y_{ij} were obtained by bootstrapping the HRF data - which means that a re-sampling with replacement was carried out in each small area (158 sampled municipalities). The population size N_i for each small area is determined by the individual's sample weights, i.e. $N_i = n_i \frac{N_t}{n_t}$. In contrast to the model-based population, the pseudo-population in this case is fixed for each simulation. The true mean for the small area *i* is calculated as: $\overline{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$. Two hundred simulations were carried out. Two hundred independently stratified random samples, with sample size n_i equal to the original sample, were selected without replacement within each small area. Each small area estimator (direct, synthetic and EBLUP), and their corresponding MSE were computed from each of the selected areas. Table 4 contains the values of *AARB*, *ARMSE*, and $AEFF_{dir}$ obtained for the direct estimator, the synthetic estimator, and the EBLUP estimator.

Table 4. Performance indicators of the three small area estimates.

Estimators	*AARB*	*ARMSE*	$AEFF_{dir}$ (%)
Direct	0.012	29.07	100
Synthetic	0.203	39.61	136.25
EBLUP	0.128	30.39	104.55

In terms of accuracy, the direct estimates (29.07) perform very similarly to the EBLUP estimates (30.39), whereas the average accuracy of the synthetic estimates increases to 39.61. According to the bias criterion (*AARB*), the EBLUP estimates have the second best performance, whereas the synthetic estimates have the worst performance (Table 4). The excellent performance of the direct estimates in this case could be attributed to the fact that reliable information can be produced using only the direct estimates for an important numbers of municipalities. In other words, the small areas with smaller coefficients of variation may have a considerable influence on the overall performance of the direct estimates over the 158 small areas. For this reason, it was decided to compare the overall estimators' performance by classifying the municipalities in three main groups: 1) 75% of the municipalities with the lowest direct *RMSE* ("*lowest 75%"*), 2) 10% of the municipalities with the highest direct *RMSE* ("*highest 10%"*), and 3) the remaining 15% of municipalities ("*75-90%"*). As a result of this, 119 out of 158 municipalities are in the *"lowest 75%"* group, 23 are

in the "*75-90%"* group, and 16 are in the "*highest 10%"* group. Table 5 gives the results obtained for each of these three groups of municipalities:

Table 5. Performance indicators for each of the three groups of municipalities.

Estimators	*AARB*			*ARMSE*			$AEFF_{dir}$ (%)		
	lowest 75%	***75-90%***	***highest 10%***	***lowest 75%***	***75-90%***	***highest 10%***	***lowest 75%***	***75-90%***	***highest 10%***
Direct	0.008	0.017	0.035	15.30	38.22	115.43	100	100	100
Synthetic	0.197	0.149	0.327	29.96	36.90	114.39	195.81	96.54	99.09
EBLUP	0.123	0.096	0.209	20.80	32.24	97.68	135.94	84.35	84.62

The Table shows that the direct estimator performs better than the synthetic and the EBLUP estimators (both in terms of bias and accuracy) for 75% of the 158 sampled municipalities (the "*lowest 75%*" group). However, for the remaining 25% of the sampled municipalities (the "*75-90%*" group, and the "*highest 10%*" group), the EBLUP estimator is better than the direct estimator, offering a gain in efficiency of about 16% (84.35% for the "*75-90%*" group, and 84.62% for the "*highest 10%*" group). The EBLUP still performs well but relative to the direct there is not a remarkable change between the two groups of municipalities, i.e. the "*75-90%*" group and the "*highest 10%*" group, because both the MSE of the EBLUP estimator and the variance of the direct estimator increase at approximately the same rate for both groups of municipalities.

5.2.1 Assesing the perfromance of the eblup mse estimators

Since the EBLUP estimator consistently performs better than the direct estimator for those municipalities with the highest direct *RMSE* (the "*75-90%*" group, and the "*highest 10%*" group) under the *ARMSE* criterion (Table 5), this section will evaluate the performance of two EBLUP MSE estimators - the mse_{PR} and the mse_{JLW} - for each of these two groups of municipalities. Table 6 presents the results of the performance of the EBLUP MSE estimators:

Table 6. Performance of each EBLUP MSE estimator.

Estimators	*AARB*		*ARB*		$AEFF_{JLW}$ **(%)**	
	75-90%	***highest 10%***	***75-90%***	***highest 10%***	***75-90%***	***highest 10%***
mse_{JLW}	0.141	0.415	0.598	-0.306	100	100
mse_{PR}	0.115	0.565	0.339	-0.535	92.61	141.92

For the "*75-90%"* group, both MSE estimators overestimate the actual MSE, although mse_{PR} overestimates less than the mse_{JLW} (11.5% for the mse_{PR} and 14.1% for the mse_{JLW}). This slight overestimation by both EBLUP MSE estimators may be because, for the vast majority of the municipalities in the "*75-90%*" group, $\hat{\sigma}_e^2$ largely overstates the actual variation in the data (see the left side of Figure 6) which may yield an overestimation of the leading term of both EBLUP MSE estimators $g_{1i}\left(\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)=\hat{\gamma}_i\left(\frac{\hat{\sigma}_e^2}{n_i}\right)$. In terms of accuracy, the mse_{PR} is slightly better than the mse_{JLW} ($AEFF_{JLW}(\%)=92.61$), leading to a gain in efficiency of about 8%. For the "*highest 10%*" group, the results are different. In terms of bias, both EBLUP MSE estimators severely underestimate the actual MSE, although the mse_{JLW} underestimates less than the mse_{PR} (Figure 6). This underestimation is of 41.5% for the mse_{JLW} and 56.5% for the mse_{PR}. This is because, for the majority of these municipalities, $\hat{\sigma}_e^2$ severely underestimates the actual variation in the data (see the right side of Figure 5). In terms of accuracy, the performance of the mse_{JLW} is much better than the performance of the mse_{PR} ($AEFF_{JLW}(\%)=141.92$).

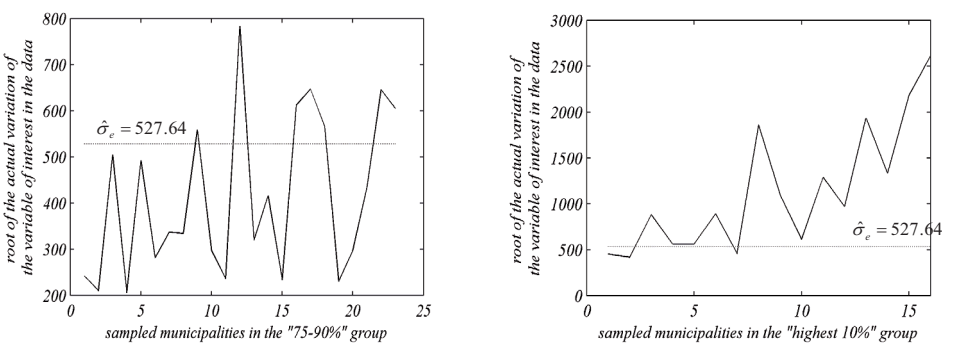

Figure 5. Root of the actual variability of the variable of interest in the data versus the root of the within area variance ($\hat{\sigma}_e$).

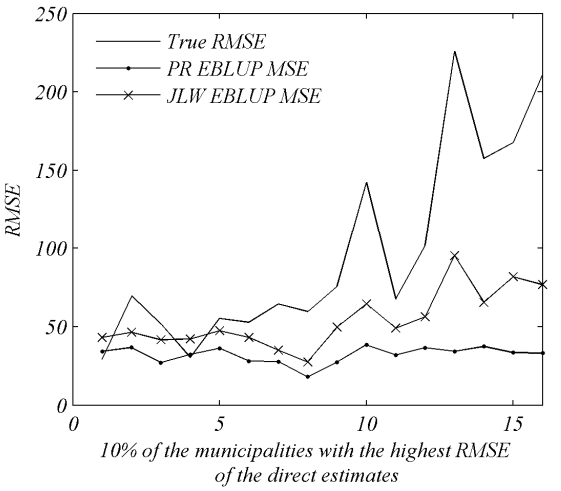

Figure 6. *RMSE* for 16 sampled municipalities in the "*highest 10%"* group.

These findings are consistent with both theoretical predictions and the simulation results as described in Jiang et al. (2002). This overestimation (underestimation) of both EBLUP MSE estimators for the municipalities in the "*75-90%*" group (the "*highest 10%*" group) is probably due to the fact that the failure of normality of the residuals causes the overestimation (underestimation) of $\hat{\sigma}_e^2$. In general, since there is a slight difference in the performance of both EBLUP MSE estimators for the municipalities that are in the "*75-90%*" group and this difference increases significantly for the municipalities that are in the "*highest 10%*" group, it is evident that the mse_{PR} are more sensitive to the non-normality of the residuals than the mse_{JLW}. Similar results were found by Fabrizi et al. (2007).

To conclude, it is suggested that the EBLUP estimates should be used to produce official statistics instead of the direct estimates for 25% of the 158 sampled municipalities (the "*75-90%*" group and the "*highest 10%*" group). Furthermore, for the 23 municipalities in the "*75-90%*" group, the mse_{PR} should be used as a measure to evaluate the quality of the information, and the mse_{JLW} should be used for the 16 municipalities in the "*highest 10%*" group. These findings may be closer to reality than the results from the model-based simulation.

6. Small area application

The core idea of this section aims to produce more precise estimates than the direct estimates of the monthly mean income at a municipal level. The 169 Cuban municipalities were considered as small areas, 158 of which were represented in the HRF sample and the remaining 11 were non-sampled municipalities. As posited in section 5 above, the design-based simulation results may be more realistic than the model-based simulation results. It was therefore decided to build the following final estimate (FE_i) for each *ith* small area (municipality):

$$FE_i = \begin{cases} \hat{\bar{Y}}_{iD} & \text{if the small area } i \text{ is in the "lowest 75\%" group} \\ \hat{\bar{Y}}_i^{EBLUP} & \text{if the small area } i \text{ is in either the "75-90\%" group or the "highest 10\%" group} \\ \hat{\bar{Y}}_{iSR} & \text{if the small area } i \text{ is a non-sampled area} \end{cases}$$

and the MSE estimator of FE_i is given by:

$$mse\left(FE_i\right) = \begin{cases} mse\left(\hat{\bar{Y}}_{iD}\right) & \text{if the small area } i \text{ is in the "lowest 75 \%" group} \\ mse_{PR}\left(\hat{\bar{Y}}_i^{EBLUP}\right) & \text{if the small area } i \text{ is in the "75 - 90 \%" group} \\ mse_{JLW}\left(\hat{\bar{Y}}_i^{EBLUP}\right) & \text{if the small area } i \text{ is in the "higest 10 \%" group} \\ mse\left(\hat{\bar{Y}}_{iSR}\right) & \text{if the small area } i \text{ is a non - sampled area} \end{cases}$$

where $\hat{\bar{Y}}_{iD}$, $\hat{\bar{Y}}_{iSR}$, and $\hat{\bar{Y}}_i^{EBLUP}$ are given in (1), (3), and (5) respectively; and $mse\left(\hat{\bar{Y}}_{iD}\right)$, $mse\left(\hat{\bar{Y}}_{iSR}\right)$, $mse_{PR}\left(\hat{\bar{Y}}_i^{EBLUP}\right)$, and $mse_{JLW}\left(\hat{\bar{Y}}_i^{EBLUP}\right)$ are described in section 3. The "*lowest 75%*" group, the "*75-90 %*" group, and the "*highest 10 %*" group were defined previously in Section 5.2. The final proposal is that the estimator, FE_i, along with its MSE estimator, $mse\left(FE_i\right)$, should be published in the future rather than the traditional direct estimator and its MSE estimator. They also permit provision of information of the monthly mean income, not only for the 158 sampled municipalities, but also for the 11 non-sampled municipalities. The gain in efficiency connected to each final municipality estimator FE_i is evaluated using the ratio of its root MSE to the root MSE of the direct estimator. Figure 7 shows the relative efficiency for the 23 municipalities that are in the "*75-*

90%" group (left side), and the 16 municipalities that are in the "*highest 10%*" group (right side). For the remaining 119 sampled municipalities in the "*lowest 75%*" group, the relative efficiency is equal to 1.

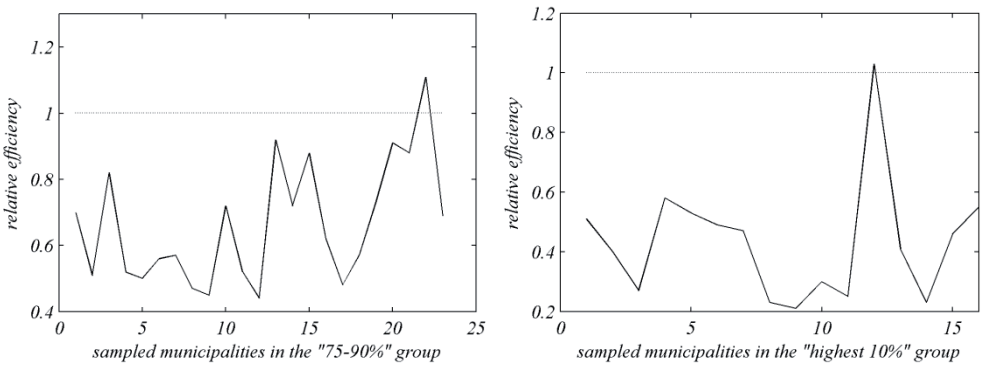

Figure 7. Relative efficiency of the "Final estimator", FE_i, with respect to the direct estimator, $\hat{\bar{Y}}_{iD}$.

The precision of the 119 sampled municipalities that are in the "*lowest 75%*" group could not be improved because the direct estimates of the monthly mean income perform acceptably well, i.e. the coefficient of variation of most of these municipalities is less than 20%. For the sampled municipalities in the "*75-90%*" group, the performance of the final estimates is observed to be better than the direct estimates in 22 out of 23 municipalities in this group. The final estimates are still better than the direct estimates even though the $mse_{PR}\left(\hat{\bar{Y}}_i^{EBLUP}\right)$, which the EBLUP MSE estimator used for these municipalities, overestimates the actual EBLUP MSE. For the 16 municipalities in the "*highest 10%*" group (right side of Figure 8), the final estimates perform better than the direct estimates in 15 of the 16, i.e. the relative efficiencies of the final estimates with respect to the direct estimates are smaller than 1. However, this assessment must be regarded with caution, considering that the design-based simulation results showed that the $mse_{JLW}\left(\hat{\bar{Y}}_i^{EBLUP}\right)$, which the EBLUP MSE estimator used for these municipalities, underestimates the actual EBLUP MSE, and therefore this relative efficiency is also underestimated. It is therefore recommended that the behaviour of other EBLUP MSE estimators should be investigated, particularly in cases where the normality assumption of the residuals does not hold. For example, different EBLUP MSE estimators can be considered to experiment with first, such as those given by the following authors: Buttar and Lahiri (2003), Pfeffermann and Glickman (2004), and Hall and Maiti (2006).

7. Final remarks and further developments

Some criticisms could be made of the results discussed in this article. The production of SAE methods depend upon two main factors, which are both addressed to find suitable models that fit the data well. The first factor relates to the quality and relevance of each of the sample data variables used to fit the model, and the second factor relates to the availability and quality of different alternative data sources, which the auxiliary information related to the target variable is selected from. With respect to the first factor, no consideration was made of the presence of outliers of the monthly total income of each individual that may affect the production of small area estimates under the basic unit level model. It is well documented that the least squares estimator of β in the model and ML or REML estimators for the variance components are sensitive to outliers (Richardson and Welsh, 1995). In order to obtain robust small area estimation with the presence of outliers we recommend analysing the feasibility of applying the methods proposed by Sinha and Rao (2009), and Chambers and Tzavidis (2006) to the HRF data. The second reason mentioned above is linked to the availability of the auxiliary information related to the variable of interest. This article chose only those variables for which municipality means were available from the 2002 Cuban census results: sex, age, education level and marital status. However, taking into account the complexity of the national Cuban monetary system (two currencies), it is advisable to keep searching for more appropriate supplementary variables – specifically economic indicators – which would improve the direct estimators, such as the proportion of workers at a small area level. It would be useful to list all the plausible variables related to income and make an inventory of the different information sources where these variables appear. The Ministry of Labour and Social Security of Cuba may be the main provider of this type of information at a small area level.

Another relevant criticism is related to the variance estimates of the direct means. In this application, the variance estimator (2) was used to estimate the precision of the direct means. This variance estimator ignores the probabilistic three-stage cluster design employed in HRF survey which may lead to a severe underestimation of the true variance. A more accurate procedure to overcome this problem could have been either to use the variance estimates obtained by applying the ultimate cluster method of variance estimation given by Hansen, Hurwitz and Madow (1953), or to develop methods to adjust the variance estimator for design effects by considering information from previous studies. A further important criticism is linked to the EBLUP MSE estimators. In this article, neither of the two EBLUP MSE estimators (mse_{PR}, and mse_{JLW}) appear to be close to the true variability when the normality assumption of the residuals does not hold. It would therefore be advisable to keep working on this topic in order to locate more robust and accurate MSE estimators. For example, as noted above, the parametric double

bootstrap technique posited by Hall and Maiti (2006) may be a possible starting point.

Despite these criticisms, this article proposes that this SAE application could constitute an important first step for extending these techniques in Cuba. The results are only meant to provide a first impression of the utility of SAE methods. More research is needed to develop a generic SAE methodology in Cuba. Finally, the implementation of the MLwiN code for this application (Appendix 1) shows the feasibility of using this statistical software in small area applications. This statistical code may be generalised to a standardized subroutine which serves to obtain different small area estimates derived from a basic unit level model and their MSE estimates.

Appendix 1. Main macro implemented in MlwiN software to calculate the two SAE methods (synthetic, EBLUP) and their corresponding MSE estimators (synthetic MSE estimator, PR EBLUP MSE estimator, and JLW EBLUP MSE estimator).

```
note running the model
resp c647
batc 1
stop
tole 2
star
note covariate matrix
obey d:/simfr/covmatrix.txt
note population mean covariate matrix
obey d:/simfr/popmatrix.txt
note sample mean covariate matrix multiply by sigma
pick 1 c1096 b1
pick 2 c1096 b2
calc c614=b1/(b1+(b2/c1105))
obey d:/simfr/samplematrix.txt
note EBLUP=EB estimates
calc c620=c1098
matr c620 21 1
calc c621=c600*c614
matr c621 158 1
calc c622=c621+(c619-c615)*.c620
note synthetic estimates
calc c623=c619*.c620
note synthetic MSE estimates
eras c599
pick 1 c1096 b1
calc c1100=sym(c1099)
obey d:/simfr/synthetic.txt
note PR EBLUP MSE estimates
note calcucale g1
calc c625=(1-c614)*b1
note calculate g2
gene 1 3318 1 c626
calc c627=c619-c615
eras c632
loop b55 1 158
calc c628=c626
obey d:/simfr/g2.txt
endl
note calculate g3
obey d:/simfr/g3.txt
note JLW EBLUP MSE estimates
put 158 0 c1051; put 158 0 c1049
loop b49 1 158
calc c5000=c1050==b49
omit 1 c5000 c5 c4999 c1020
omit 1 c5000 c14 c4999 c1021
omit 1 c5000 c15 c4999 c1022
omit 1 c5000 c18 c4999 c1023
omit 1 c5000 c21 c4999 c1024
omit 1 c5000 c22 c4999 c1025
omit 1 c5000 c23 c4999 c1026
omit 1 c5000 c26 c4999 c1027
omit 1 c5000 c27 c4999 c1028
omit 1 c5000 c28 c4999 c1029
omit 1 c5000 c29 c4999 c1030
omit 1 c5000 c30 c4999 c1031
omit 1 c5000 c31 c4999 c1032
omit 1 c5000 c32 c4999 c1033
omit 1 c5000 c33 c4999 c1034
omit 1 c5000 c34 c4999 c1035
omit 1 c5000 c35 c4999 c1036
omit 1 c5000 c36 c4999 c1037
omit 1 c5000 c37 c4999 c1038
omit 1 c5000 c38 c4999 c1039
omit 1 c5000 c39 c4999 c1040
omit 1 c5000 c1050 c4999 c1043
pick b49 c1105 b1
calc b2=22851-b1
put b2 1 c1041
gene 1 b2 1 c1042
note build the model with the deleted observations
clea
resp c1020
iden 2 c1043
iden 1 c1042
expl 1 c1041
setv 2 c1041
setv 1 c1041
```

```
note calc PR EBLUP MSE estimates = g1+g2+2*g3
calc c634=c625+c632+2*c633
```

```
addt 'c1021'; addt 'c1022'; addt 'c1023';
addt 'c1024'; addt 'c1025'; addt 'c1026';
addt 'c1027'; addt 'c1028'; addt 'c1029';
addt 'c1030'; addt 'c1031'; addt 'c1032';
addt 'c1033'; addt 'c1034'; addt 'c1035';
addt 'c1036'; addt 'c1037'; addt 'c1038';
addt 'c1039'; addt 'c1040'
note running the model
tole 2
meth 0
batc 1
star
pick 1 c1096 b3
pick 2 c1096 b4
calc c1044=b3/(b3+(b4/c1105))
calc c1045=(1-c1044)*b3
calc c1051=c1051+(c1045-c625)
calc c614=c1044
obey d:/simfr/samplematrix.txt
calc c1046=c1098
matr c1046 21 1
calc c1047=c600*c614
matr c1047 158 1
calc c1048=c1047+(c619-c615)*.c1046
calc c1049=c1049+(c1048-c622)^2
endl
note JLW EBLUP MSE estimates
calc c1051=(157/158)*c1051
calc c1049=(157/158)*c1049
calc c1053=c625-c1051+c1049
```

REFERENCES

BUTTAR, F., AND LAHIRI, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference,* 112, 63-76.

CHAMBERS, R., and TZAVIDIS, N. (2006). "M-quantile models for small area estimation". *Biometrika* 93(2):255-268.

FABRIZI, E., FERRANTE, M.R., and PACEI, S. (2007). Small area estimation of average household income based on unit level models for panel data. *Survey Methodology*, vol. 33, No. 2, pp. 187-198. Statistics Canada, Catalogue no.12-001-X.

GOLDRING, S., LONGHURST, J., and CRUDDAS, M. (2005). Model-Based Estimates of Income for wards, 2001/2002. Technical Report. *Office for National Statistics* (ONS).

GOLDSTEIN, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76: 622-623.

HALL, P., and MAITI, T. (2005). Nonparametric estimation of mean-squared prediction error in nested-error regression models. Sep. http://arxiv.org/abs/math/0509493.

HANSEN, M.H., HURWITZ, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory. *New York, Wiley*.

JIANG, J., LAHIRI, P., AND WAN, S.M. (2002). A unified jackknife theory for empirical best prediction with *M*-estimation. *The Annals of Statistics*, 30, 1782-1810.

ONE. (2005). Diseño Muestral General del sistema de encuesta de hogares." Havana. *Oficina Nacional de Estadísticas*.

PFEFFERMAN, D., AND GLICKMAN, H. (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. *ASA Proceedings of the Survey Research Methods Section.*

PRASAD, N.G.N., AND RAO, J.N.K (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, J.N.K. (2003). Small area estimation. *Wiley series in survey methodology*.

RICHARDSON, A.M., and WELSH, A.H. (1995). Robust Restricted Maximum Likelihood in Mixed Linear Models. *Biometrics*, 51, 1429-1439.

SARNDAL, C., SWENSSOn B., and WRETMAN J. (1992). Model assisted survey sampling. *Springer Series in Statistics*.

SINHA, S.K., and RAO, J.N.K. (2009). Robust Small Area Estimation. *The Canadian Journal of Statistics*, to appear.