ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 8, 2004

Comparison of small area estimation methods: simulation study in EURAREA project

Kaja Sõstra

Abstract. In literature several techniques have been introduced for small area estimation. It has not been proved theoretically that some estimator is superior to others. The focus of this paper is on comparison of some estimators (direct, GREG and EBLUP estimators with correlated area and time effects) in a simulation study. The individual-level models are used. The target of estimation is the average disposable income in small geographical regions of Finland. The results given here are developed in the framework of EURAREA project, where the author of this paper was responsible on the simulation study.

1. Introduction

Small area estimation is becoming important in survey sampling due to growing demand for reliable small area statistics. Data obtained from sample surveys can be used to derive reliable direct estimates for large areas or domains, but sample sizes in small areas are rarely large enough for direct estimators to provide adequate precision for small areas. This makes it necessary to "borrow strength" from related areas to find indirect estimators that increase the effective sample size and thus increase the estimation precision.

EURAREA (Enhancing Small Area Estimation Techniques to Meet European Needs) is a three year project, funded by the European Community, which is being undertaken by a consortium of 6 European National Statistical Institutes (NSIs) and 5 universities, covering altogether 7 European countries. The overall aim of the project is to improve small area estimation methods currently used within European NSIs.

Received November 11, 2003.

2000 *Mathematics Subject Classification.* 62D05, 62H12, 62J05.

Key words and phrases. Small area estimation, linear mixed model, spatial correlated area effects, correlated time effects.

The first part of the project consists of testing of standard small area estimation techniques involving eight estimators. Among them direct, GREG and EBLUP estimators are discussed in this paper. Secondly the enhanced estimators borrowing strength over time and space were developed in the project. From them the EBLUP estimators with correlated spatial and time effects are considered in this paper. Comparison of different techniques was based on real data to stay as close as possible to the real life situation. The sample designs used to collect the data were taken into account.

Theoretically it is not proved which small area estimator is the best. Estimators perform differently depending on several conditions. National surveys take place on the same circumstances during long time – the same set of target variables, the same sub-populations (incl. small areas), the same sampling design. Consequently it is possible to organise simulation study close to the real situation for testing performance of small area estimators for a particular survey and to suggest better performing small area estimators for use in practice.

Large simulation study was designed and conducted in Statistics Finland in the framework of EURAREA project to test small area estimators for average disposable income. For simulations the register-based employment statistics database of Finnish population was used. The database includes about 5.7 million records with employment and other characteristics covering time period 1987-1998.

2. Estimators

The following five small area estimators are studied in this paper:
1. Direct estimator;
2. GREG estimator with a standard linear regression model;
3. The three EBLUP estimators based on linear two-level models with individual-level data assuming
   (a) independent spatial effects denoted by EBLUP();
   (b) correlated spatial effects EBLUP(S);
   (a) independent spatial effects and correlated time effects EBLUP(T).

2.1. Direct estimator. The direct estimator of the mean in the area d is defined as a ratio of the design-weighted Horvitz–Thompson estimators for each area (EURAREA, 2004, p. 13):

$$\widehat{\overline{Y}}_d = \sum_{i \in s_d} w_i \, y_i / \widehat{N}_d \ ,$$

where

$$\widehat{N}_d = \sum_{i \in s_d} w_i \ .$$

The sums are taken over sample s_d from area d and design weights are inverses of inclusion probabilities, $w_i = \pi_i^{-1}$. The precision of the estimator is measured by its mean square error, estimated by the following formula (Särndal et al, 1992, p. 391):

$$M\widehat{SE}\left(\widehat{\overline{Y}}_d\right) = \sum_{i\in s_d}\sum_{j\in s_d} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\,\frac{\left(y_i-\widehat{\overline{Y}}_d\right)}{\pi_i}\,\frac{\left(y_j-\widehat{\overline{Y}}_d\right)}{\pi_j}\Big/\widehat{N}_d^2.$$

Assuming independence, $\pi_{ij} = \pi_i \cdot \pi_j$, whenever $i \neq j$ we get:

$$M\widehat{SE}\left(\widehat{\overline{Y}}_d\right) = \sum_{i\in s_d} w_i\ (w_i - 1)\ \left(y_i - \widehat{\overline{Y}}_d\right)^2 \Big/ \widehat{N}_d^2.$$

2.2. GREG estimator. The generalised regression estimator (GREG) is obtained by adjusting the direct estimator using the standard linear model. The formula for GREG estimator is (EURAREA, 2004, p. 14):

$$\widehat{\overline{Y}}_d^{GREG} = \widehat{\overline{Y}}_d + \left(\overline{\mathbf{X}}_d - \widehat{\overline{\mathbf{X}}}_d\right)^T \widehat{\boldsymbol{\beta}},$$

where

$$\widehat{\overline{\mathbf{X}}}_d = \sum_{i\in s_d} w_i\ \mathbf{x}_i / \widehat{N}_d\ ,$$

$\overline{\mathbf{X}}_d = \left(\overline{X}_{d,1}, ..., \overline{X}_{d,p}\right)^T$ is the vector of true means of p covariates ($\mathbf{x}_i$ is p-dimensional) in the area d and $\widehat{\boldsymbol{\beta}}$ is the least squares regression estimate assuming a standard linear model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$ with independent errors $\epsilon_i \sim N(0, \sigma^2)$ for each unit i in the sample:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i\in s} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i\in s} w_i \mathbf{x}_i y_i.$$

Note that β is estimated using the whole sample s. An alternative presentation for GREG estimator is given through g-weights:

$$\widehat{Y}_d^{GREG} = \sum_{i\in s} w_i g_{di} y_i,$$

where g-weights depend on the domain d, element i and the whole sample s:

$$g_{di} = \frac{N_d}{\widehat{N}_d} z_{di} + N_d\left(\overline{\mathbf{X}}_d - \widehat{\overline{\mathbf{X}}}_d\right)^T \left(\sum_{i\in s} w_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \mathbf{x}_i,$$

with domain indicators $z_{di} = 1$, if $i \in s_d$ and $z_{di} = 0$, otherwise. Elements of the full sample s that belong to domain d form the domain sample s_d.

An estimate of the mean square error of GREG estimator is derived as follows (Särndal et al, 1992, p. 401):

$$M\widehat{S}E\left(\widehat{\overline{Y}}_d\right) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{g_{di}r_i}{\pi_i}\frac{g_{dj}r_j}{\pi_j}/\widehat{N}_d^2,$$

with residuals $r_i = y_i - \mathbf{x}_i^T\,\widehat{\boldsymbol{\beta}}$. Assuming that $\pi_{ij} = \pi_i \cdot \pi_j$, whenever $i \neq j$, we get:

$$M\widehat{S}E\left(\widehat{\overline{Y}}_d^{GREG}\right) = \sum_{i\in s} w_i\ (w_i - 1)\ g_{di}^2\ r_i^2/\widehat{N}_d^2.$$

2.3. EBLUP estimator with independent spatial effects. Let the population model of interest be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is the study variable with values for each unit in the population, $\mathbf{X}$ is the matrix of auxiliary information with dimension $N \times p$, β is the p-dimensional vector of regression coefficients and $\mathbf{Z}$ is $N \times D$ incidence matrix for D-dimensional area random effect vector $\mathbf{u}$.

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{N_1} & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \mathbf{1}_{N_D} \end{bmatrix},$$

where $\mathbf{1}_{N_d}$ is a vector of dimension N_d with all elements equal to one.

Vectors $\mathbf{e}$ and $\mathbf{u}$ follow multinormal distribution with zero mean vectors and covariance matrices of $\mathbf{G} = \sigma_e^2\mathbf{I}_N$ and $\mathbf{R} = \sigma_u^2\mathbf{I}_D$ respectively. Matrices $\mathbf{I}_N$ and $\mathbf{I}_D$ are the identity matrices with dimensions N and D respectively. The covariance matrix of $\mathbf{y}$ is $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T$ (Rao, 2003, p. 96).

We assume that model (1) holds in the sample. Let s refer to the sampled part and r to the non-sampled part of the population. Our aim is to estimate a linear function

$$\boldsymbol{\theta} = \mathbf{L}\mathbf{y} = \mathbf{L}_s\mathbf{y}_s + \mathbf{L}_r\mathbf{y}_r. \tag{2}$$

The first term of (2) is known when survey is conducted, the second term can be predicted:

$$\widehat{\boldsymbol{\theta}} = \mathbf{L}_s\mathbf{y}_s + \mathbf{L}_r\left(\mathbf{X}_r\widehat{\boldsymbol{\beta}} + \mathbf{Z}_r\widehat{\mathbf{u}}\right). \tag{3}$$

Assuming known variance components $\mathbf{G}$ and $\mathbf{R}$, the best linear unbiased predictor (BLUP) method provides estimators of $\boldsymbol{\beta}$ and $\mathbf{u}$ for the second term of (2) (Rao, 2003, p. 96):

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y},$$

$$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

In practice we need to estimate variance components. The empirical best unbiased predictor (EBLUP) replaces the unknown variance components in the BLUP by these estimates.

2.4. EBLUP estimator with correlated spatial effects. Theory for EBLUP estimators with correlated area and time effects is introduced in Chambers and Saei (2003a) and Chambers and Saei (2003b). The area effects $\mathbf{u}$ introduced in the previous section are independent. In real life situations area effects are often correlated. The value of the correlation coefficient may depend on many different factors. In this paper it is assumed that the distances between the geographical central points of regions affect correlations.

Now the vector $\mathbf{u}$ in the model (1) follows multinormal distribution with zero mean vectors and covariance matrices $\sigma_u^2\mathbf{A}$ where the correlation matrix $\mathbf{A}$ takes a form,

$$\mathbf{A} = [a_{is}] = [\exp(-\alpha \cdot |d_i - d_s|)],$$

where α is an unknown parameter and $|d_i - d_s|$ is the distance between areas i and s.

The predictors $\widehat{\boldsymbol{\theta}}$ can be estimated using similar methods as described in the previous section.

2.5. EBLUP estimator with correlated time effects. Many surveys conducted by national statistical institutes are continuous. This means that past values of the variable of interest can be used for "borrowing strength over time". Let vector $\mathbf{u}_1$ represent time effect and vector $\mathbf{u}_2$ area effect. The population model takes the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e}. \tag{4}$$

Vectors $\mathbf{e}$, $\mathbf{u}_1$ and $\mathbf{u}_2$ follow multinormal distribution with zero mean vectors and covariance matrices $\sigma_e^2\mathbf{I}_N$, $\sigma_{u1}^2\mathbf{A}_1$ and $\sigma_{u1}^2\mathbf{I}_D$ respectively. Assuming the first-order autoregressive model for $\mathbf{u}_1$ with T time-points the matrix $\mathbf{A}_1$ is

$$\mathbf{A}_1 = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \dots & \rho^{T-2} \\ \dots & \dots & \dots & \dots \\ \rho^{T-1} & \rho^{T-2} & \dots & 1 \end{bmatrix}.$$

Setting $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2')'$ and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ model (4) simplifies to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}.$$

The predictors $\widehat{\boldsymbol{\theta}}$ can be calculated by (3).

3. Simulation study

Finnish register-based employment statistics database was used for a simulation study. Database includes about 5.7 millions records (2.0 millions in Western Finland) for time period between 1987 and 1998. Population for the simulation study was Western Finland data in 1998 (first four estimators) and between 1994-1998 (EBLUP estimator with correlated time effects). Every record consists of about 60 variables per year including personal information (age, sex, education, language, marital status), household type information (dwelling, place of residence), data about working life, income and taxes. Software for simulations was worked out in the Office of National Statistics (UK) and Statistics Finland.

3.1. Sample design and data issues. 1000 samples were drawn from the database by stratified simple random sampling for time period 1994-1998. The overall sample size was 10,000 households (2000 households per year). Table 3.1 shows stratification of the sample by socio-economic groups. Farmers and other entrepreneurs with unstable income were oversampled to get the sample design close to the real design used in national surveys in Statistics Finland.

Table 3.1. Stratification by socio-economic groups

Strata	n	incl. prob.
wage and salary earners	830	0.00138
farmers	270	0.00618
other entrepreneurs	270	0.00448
pensioners	330	0.00086
other socio-economic categories	250	0.00083
not specified (mainly children)	50	0.00063
sample size total	2000	0.00227

Table 3.2. Number of sub-regions by sample size in Western Finland

Sample size	Number of sub-regions
12-19	5
20-29	10
30-39	6
40-49	4
50-100	7
120-284	4
Small (<40)	21
Large (>40)	15
Total	36

There are 21 regions with relatively small sample size in Western Finland (see table 3.2). All simulations results are analysed in two groups: large regions with average sample size over 40 and small regions with average sample size less than 40 households.

For each household the disposable income (DI) was defined:

$$DI = \frac{\text{Total Household Net Income}}{1 + 0.5((\#\text{ persons aged } \geq 14) - 1) + 0.3(\#\text{ children aged} < 14)}.$$

Average disposable income was estimated in Western Finland on the subregional level.

All estimators (except direct one which does not use any auxiliary information) use the same two auxiliary variables:

- the number of persons having tertiary education in the household;

- the total number of months of all household members being in employment per year.

3.2. Simulation results. Two measures were applied to compare the performance of the different estimators for M=1000 simulations, the mean absolute relative bias

$$MARB = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{M}\sum_{m=1}^{M}\frac{\widehat{\overline{Y}_d}^{(m)} - \overline{Y}_d}{\overline{Y}_d}\right|$$

and the relative root mean square error

$$RRMSE = \frac{1}{D}\sum_{d=1}^{D}\frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\overline{Y}_d}^{(m)} - \overline{Y}_d\right)^2}}{\overline{Y}_d},$$

where D refers to the number of regions, $\widehat{\overline{Y}_d}^{(m)}$ is the predicted value of the average disposable income from mth simulation in region d and $\overline{Y}_d$ refers to the true population mean of the disposable income in the same region.

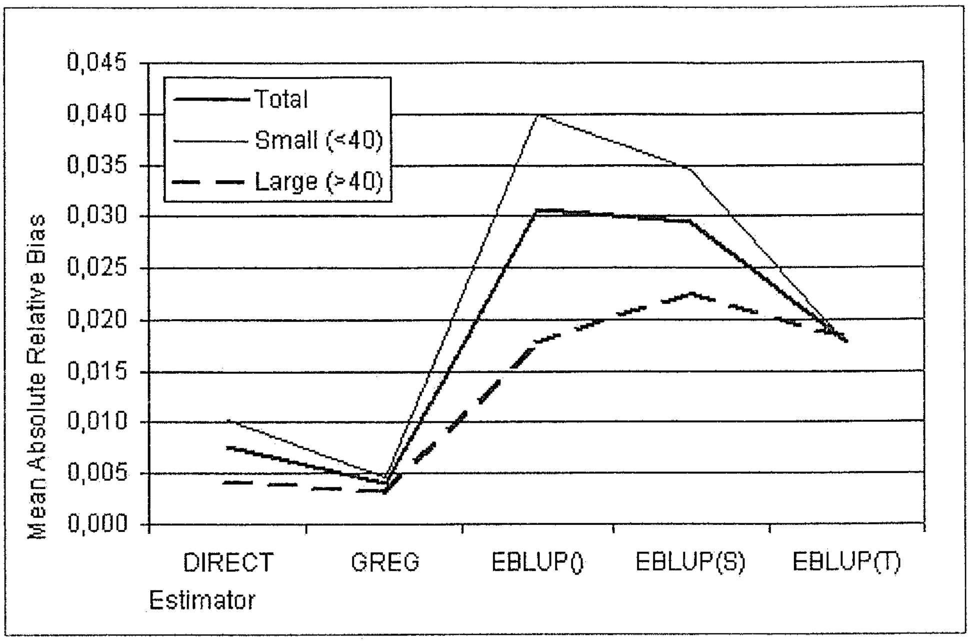

Figure 3.1. Mean absolute relative bias

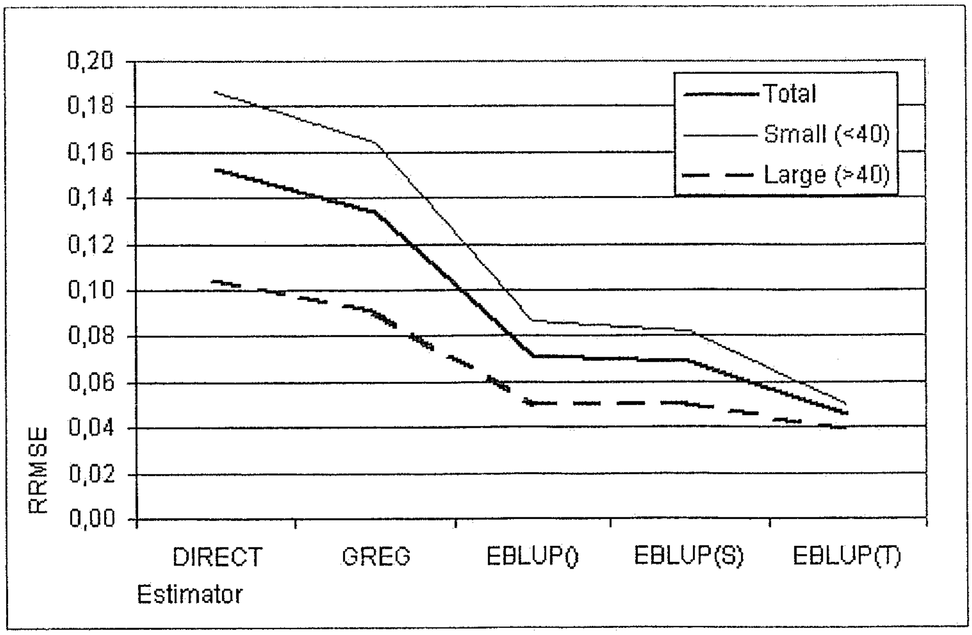

Figure 3.2. Relative root mean square error

Figures 3.1 and 3.2 show the performance of five estimators: direct, GREG, EBLUP(), EBLUP(S), EBLUP(T).

With respect to the bias measure direct and GREG estimators perform better than others. Adding spatial information to the model improves slightly performance of EBLUP estimator in smaller regions. Using time series information decreases bias of EBLUP estimator considerably (Figure 3.1).

According to the relative root mean square error (RRMSE) all EBLUP estimators outperform the direct and GREG (Figure 3.2). Best result is obtained with EBLUP(T), especially in smaller regions.

4. Conclusions

The comparison of the five estimators revealed:

1) Precision of direct and GREG estimators is very low compared to model-based EBLUP estimators.

2) EBLUP estimator with correlated spatial effects performs almost as well as EBLUP estimator with independent spatial effects.

3) EBLUP estimator with correlated time effects outperforms other EBLUP estimators significantly.

Simulation results suggest to use maximum existing information about the target variable. The first important step is the choice of the best possible set of auxiliary variables. The correlation between auxiliary and target variables was quite low in this simulation study (<0.2) but we still experienced improvement of GREG estimator which uses auxiliary information compared to the direct estimator.

The second step is the choice of appropriate estimator. Direct estimator is quite unstable if sample size is relatively small. GREG estimator gave a slight improvement compared to the direct one. Good results are obtained with GREG estimator if strong auxiliary information is available, but for areas with very small sample size GREG estimator gives unstable results. Direct and GREG estimators cannot estimate small areas with zero sample size. EBLUP estimators are recommended if sample sizes are very small or zero for some areas. The best results are received by "borrowing strength" over time. Using available time series information is always recommendable. EBLUP estimator with correlated time effects has lower bias than other model-based estimators and significantly higher precision compared to all other estimators observed in this paper.

Acknowledgements. The author is thankful to EURAREA project IST 2000-26290 under which most of the material of this paper was worked out, and also for the support of Estonian Science Foundation Grant 5523 and of Target Financed Project 0181776s01.

References

Chambers, R. and Saei, A. (2003a). *Linear Mixed Model with Spatial Correlated Area Effect in Small Area Estimation. EURAREA Project.* Department of Social Statistics, University of Southampton, United Kingdom.

Chambers, R. and Saei, A. (2003b). *Small Area Estimation Based on a Unit Level Linear Mixed Model with Correlated Time Effect. EURAREA Project.* Department of Social Statistics, University of Southampton, United Kingdom.

EURAREA (2004). *Report on the Performance on "Standard" Estimators.* The EURAREA Consortium, Office for National Statistics, United Kingdom.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley, New York.

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

Statistical Office of Estonia, Vabaduse pl 4, 71020 Viljandi, Estonia; Institute of Mathematical Statistics, J. Liivi 2, 50409 Tartu, Estonia

E-mail address: kaja.sostra@stat.ee