Liet. matem. rink., T. 43, spec. nr., 497–503

An example of small area estimation in finite population sampling

Danutė KRAPAVICKAITĖ (MII)
e-mail: krapav@ktl.mii.lt

1. Introduction

The estimator of a parameter in some area of population is called a small area estimator if it uses data from the neighboring areas.

The need for the estimator of this kind arises where the sample size in the area of estimation is small and the accuracy of a direct estimator, based only on the sample data of this area, is not sufficient or sometimes there are no observations in this area at all.

Various models are often used in order to improve precision of the estimators. Auxiliary data are needed for this. The data available depend on the country. It is impossible to say in advance which estimator is the most suitable one in each situation. Experimental estimation is needed. This kind of experiment is done in this paper for the income per capita in the rural area of Lithuania.

Two kinds of small area estimators – James–Stein estimator ([2]) and empirical best linear unbiased predictor (EBLUP) ([3]) – are used for the data of the Lithuanian household budget survey (HBS). The direct estimates, based on the sample design, and currently used calibrated estimates are also presented to compare.

The results of the paper are aimed at choosing the most suitable estimators in HBS and can be helpful for similar purposes in other surveys, too.

2. Household budget survey

HBS is one of the most important sample surveys in official statistics of every country. It estimates income in cash and kind and expenditure per capita of the population of the country and in various parts of the population. Sometimes these estimates have inadmissibly high variances.

The Lithuanian HBS uses a stratified sampling design with one stage sampling in the cities and two stage cluster sampling in the medium and small towns and the rural area. Administrative division of Lithuania consists of 10 counties which are divided into districts. When sampling in the rural area, a sample of districts is drawn in the first stage and households are selected in the second stage. The sample size in the district is 10–30 households in a quarter, and part of the districts is not sampled at all.

Let us denote by $\mathcal{U}$ the finite population of size N. $\mathcal{U}_i, \mathcal{U}_i \subset \mathcal{U}$, are population areas of sizes N_i and sample sizes in those areas $n_i,\ i = 1, \ldots, m, n_1 + \ldots + n_m = n$. Let us

denote by y the study variable with values y_{ij}, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$ in the areas. If y denotes income of a person living in the rural area, income per capita in the area $\mathcal{U}_i$ is a ratio $I_i = t_{yi}/N_i$, $t_{yi} = \sum_{j \in \mathcal{U}_i} y_{ij}$. Its direct, design based estimator $\widehat{I}_i = \widehat{t}_{yi}/\widehat{N}_i$ is used.

The coefficient of variation of the estimates is lower than 3% in most cases, and it reaches 18% in some cases. The accuracy of the estimates in the areas is supposed to be not sufficiently high because of the too small sample size in those areas.

Let us denote the parameter to be estimated (a function of totals) by θ_i, $i = 1, \ldots, m$, and its direct estimator by $\widehat{\theta}_i$.

3. Methods used

3.1. *Standard linear regression model*

Suppose there are k auxiliary variables $x_1, \ldots, x_k$ with values x_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, m$, known on the area level, characterizing those areas. Then the standard linear regression model for the direct estimators $\widehat{\theta}$ can be used to predict θ:

$$\widehat{\theta}_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + u_i, \tag{1}$$

$i = 1, \ldots, m$ with the model parameters $\beta_0, \beta_1, \ldots, \beta_k$ and independent random errors u_i with normal distribution $N(0, \sigma^2)$.

The ordinary least squares (OLS) estimation of the model parameters $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ yields us an estimator

$$\widehat{\theta}_i^0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \ldots + \widehat{\beta}_k x_{ki}. \tag{2}$$

It is called a synthetic estimator: $\widehat{\theta}_i^0 = \widehat{\theta}_i^0(synth)$. It allows us to predict the parameter in any area with the known values of the auxiliary variables. It is possibly biased.

3.2. *James–Stein estimator*

Assume the direct estimators $\widehat{\theta}_i$, $i = 1, \ldots, m$ to be independent normally distributed random variables with means θ_i and a common known variance ψ. The James–Stein estimator (Fay *et al.*, [2]) is given by

$$\widehat{\theta}_i(JS) = \widehat{\phi}_{JS}\widehat{\theta}_i + (1 - \widehat{\phi}_{JS})\widehat{\theta}_i^0, \quad i = 1, \ldots, m, \tag{3}$$

here

$$1 - \widehat{\phi}_{JS} = \frac{\psi}{d/(m - 2 - k)}, \quad m > 3$$

with $d = \sum_{i=1}^{m} (\widehat{\theta}_i - \widehat{\theta}_i^0)^2$ and the standard linear regression model $\widehat{\theta}_i^0$ (2), based on k regressors. The James–Stein estimator belongs to the class of composite estimators.

This estimator shifts the direct estimator $\widehat{\theta}_i$ in the direction of the model based estimator $\widehat{\theta}_i^0$.

A direct estimator is usually unbiased or approximately unbiased and has high variance. A synthetic estimator is usually biased and has small variance. Composite estimators, like the James–Stein estimator, are a compromise between the latter two.

The variance ψ is actually unknown, and it is estimated by

$$\widehat{\psi} = \frac{1}{m}\sum_{l=1}^{m}\frac{1}{n_l}\frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2,$$

$\bar{y}_i$ is the sample average in the area i. Nevertheless, the fact of estimation of ψ is not taken into account when estimating $MSE(\theta_i(JS))$.

The mean square error of the James–Stein estimator equals

$$MSE(\widehat{\theta}_{yi}(JS)) \approx \widehat{\phi}_i^2 Var\,\widehat{\theta}_i + (1-\widehat{\phi}_i)^2 MSE(\widehat{\theta}_i^0), \quad i = 1, 2, \ldots, m$$

with $\widehat{MSE}(\widehat{\theta}_i^0) = (\widehat{\theta}_i^0 - \widehat{\theta}_i)^2$. It is unstable. The average over the small areas is also used as an estimator of $MSE(\widehat{\theta}_i(JS))$ and it is more stable:

$$\widehat{MSE}(JS) = \frac{1}{m}\sum_{i=1}^{m}\widehat{MSE}(\widehat{\theta}_i(JS)).$$

3.3. *Empirical best linear unbiased predictor*

The weights in expression (3) of the James–Stein estimator are constant for any area. If the coefficients depend on the area, it can be expected to get a higher precision of the estimator. EBLUP is on of the estimators of such kind.

We assume that the direct estimators $\widehat{\theta}_i$ can be expressed by

$$\widehat{\theta}_i = \theta_i + e_i, \quad i = 1, \ldots, m, \tag{4}$$

here e_i's are independent sampling errors with distributions $N(0, \psi_i)$ for fixed θ_i. We assume also the superpopulation model

$$\theta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + v_i, \tag{5}$$

here v_i's are independent and identically distributed random variables with normal distribution $N(0, \sigma_v^2)$. Combining (4) and (5), we obtain the area level model

$$\widehat{\theta}_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \xi_i, \tag{6}$$

here $\xi_i = v_i + e_i$ are independent random variables with the distribution $N\big(0, \sigma_v^2 + \psi_i\big)$. The variances ψ_i, $i = 1, \ldots, m$ are unknown, and can be estimated by $\widehat{\psi}_i = \widehat{\sigma}_e^2/n_i$,

$$\widehat{\sigma}_e^2 = \frac{1}{n-m}\sum_{i=1}^{m}\sum_{j}(y_{ij}-\bar{y}_i)^2,$$

here n is the number of elements in the sample, $\bar{y}_i$ is the sample average of y in the area i. Estimation of this regression model gives us the synthetic estimator of θ

$$\widehat{\theta}_i^{(s)} = \mathbf{x}_i^T \widetilde{\beta}(\widehat{\sigma}_v^2), \tag{7}$$

with

$$\widetilde{\beta}(\widehat{\sigma}_v^2) = \left(\mathbf{X}^T \mathbf{D}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{D}^{-1} \widehat{\theta}.$$

$\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)^T$, $\mathbf{X}$ is a matrix, the columns of which are the vectors $\mathbf{x}_i = (1, x_{1i}, \ldots, x_{ki})^T$. $\mathbf{D}$ is a diagonal matrix with $\left(\widehat{\sigma}_v + \frac{\widehat{\sigma}_e^2}{n_i}\right)$ on the diagonal, $\widehat{\sigma}_v = \max(0, \widetilde{\sigma}_v)$,

$$\widetilde{\sigma}_v = \frac{1}{m-k}\Big(\sum_{i=1}^{m} (\widehat{\theta}_i - \mathbf{x}_i^T \widehat{\beta})^2 - \sum_{i=1}^{m} \frac{\widehat{\sigma}_e^2}{n_i} \big(1 - \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i\big)\Big),$$

$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k)^T$ is the estimator of standard regression coefficients (2). The mean square error of the estimator (7) can be estimated by

$$\widehat{MSE}(\widehat{\theta}_i^{(s)}) = \widehat{\sigma}_v^2 + \mathbf{x}_i^T \widehat{\mathbf{V}} \mathbf{x}_i, \quad \widehat{\mathbf{V}} = (\mathbf{X}^T \mathbf{D}^{-1} \mathbf{X})^{-1}.$$

The variances ψ_i are considered here as known, and their estimation is not taken into account in $\widehat{MSE}(\widehat{\theta}_i^{(s)})$.

In order to avoid a possible bias of estimator (7), the composite estimator (EBLUP) is used with area specific weights, in contrast to the James–Stein estimator (3), Ghosh *et al.* [3]:

$$\widehat{\theta}_i^{EB}(\widehat{\sigma}_v^2) = \widehat{\gamma}_i \widehat{\theta}_i + (1 - \widehat{\gamma}_i) \mathbf{X}_i^T \widetilde{\beta}(\widehat{\sigma}_v^2) \tag{8}$$

with

$$\widehat{\gamma}_i = \frac{\widehat{\sigma}_v^2}{\widehat{\sigma}_v^2 + \widehat{\sigma}_e^2 / n_i}.$$

The estimate of its mean square error is

$$\begin{aligned}
&\widehat{MSE}(\widehat{\theta}_i^{EB}(\widehat{\sigma}_v^2)) = g_{1i}(\widehat{\sigma}_v^2) + g_{2i}(\widehat{\sigma}_v^2) + g_{3i}(\widehat{\sigma}_v^2),\\
&g_{1i}(\widehat{\sigma}_v^2) = \widehat{\gamma}_i \widehat{\sigma}_e^2 / n_i, \quad g_{2i}(\widehat{\sigma}_v^2) = (1 - \widehat{\gamma}_i)^2 \mathbf{x}_i^T (\mathbf{X}^T \mathbf{D}^{-1} \mathbf{X})^{-1} \mathbf{x}_i,\\
&g_{3i}(\widehat{\sigma}_v^2) = \frac{(\widehat{\sigma}_e / n_i)^2}{(\widehat{\sigma}_v^2 + \widehat{\sigma}_e^2 / n_i)^3} \frac{2}{m^2} \sum_{j=1}^{m} \left(\widehat{\sigma}_v^2 + \frac{\widehat{\sigma}_e^2}{n_j}\right)^2, \quad j = 1, \ldots, m.
\end{aligned}$$

4. Estimation results

The data of the HBS survey of the fourth quarter of 2002 are used here for estimation. Demographical data and agricultural production data are used to build an income model. It has been found that the variables significantly affecting income per capita are:

MILK – amount of milk produced per citizen in the rural area of the district,

CORN – amount of corn yield produced per citizen in the rural area of the district,

MEN50_65 – part of citizens of the rural area of the district composed of men aged 50–65,

WOMEN50 – part of citizens of the rural area of the district consisting of women over 50,

EMP – employment level in the district,

SOC – social allowances per citizen in the district.

The model obtained is presented in Table 1.

All the explanatory variables are significant in the model except the employment level. 63% of the total sum of squares is explained by the model. The highest correlation between $\widehat{\theta}$ and auxiliary variables is $corr(\widehat{\theta}, \log(SOC)) = -0.44$, other correlations are very low.

The estimates are presented in Fig. 1. The domains of the estimation (districts) are ordered by the sample size (number of individuals), and vary from 33 to 92. The estimates of the mean square errors of the estimators for districts, are presented in Fig. 2.

Six different estimates are represented. *I direct* is a direct estimate of income based on the sample design, no auxiliary information used. *I OLS* is an ordinary least squares regression model, described in Table 1, the sample design is avoided. *I JS* is the James–Stein estimator with *I OLS* as a synthetic part. *I sint* – regression model (6) is used. *I EBLUP* is the best linear unbiased predictor (8) used for the estimation. *I cal* is a calibrated estimator, currently used in the real survey. Calibration of design weights is used in order to adjust the sample to the demographic data and to nonresponse (Deville *et al.* [1]).

The estimated weight in the James–Stein estimator (3) is $\widehat{\phi}_{JS} = 0.64$, the average weight in the EBLUP estimator $\bar{\gamma} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\gamma}_i = 0.6$ and does not differ significantly from that of $\widehat{\phi}_{JS}$.

Table 1

Income model

Auxiliary variable	$\widehat{\beta}$	t statistic	p value
Intercept	387	2,62	0.0030
CORN and MEN50–65	3302	3,55	0.0019
CORN and MILK	−187	−3,36	0.0030
log(MILK)	−151	−4,29	0.0003
MILK and WOMEN50	1204	4,44	0.0002
EMP	−156	−1,51	0.1453
log(SOC)	−79	−2,34	0.0292

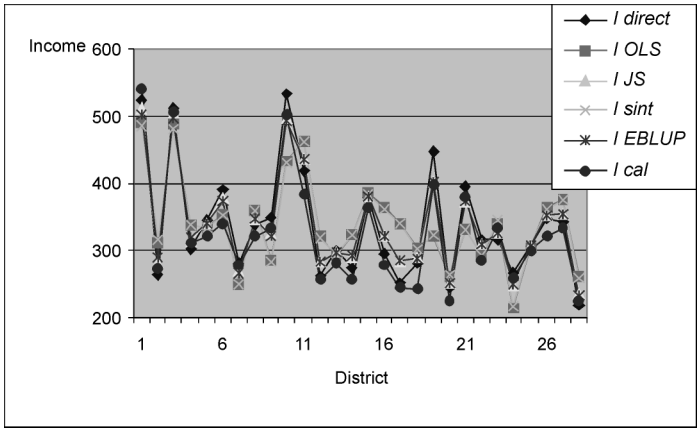

Fig. 1. Estimates of the income per capita.

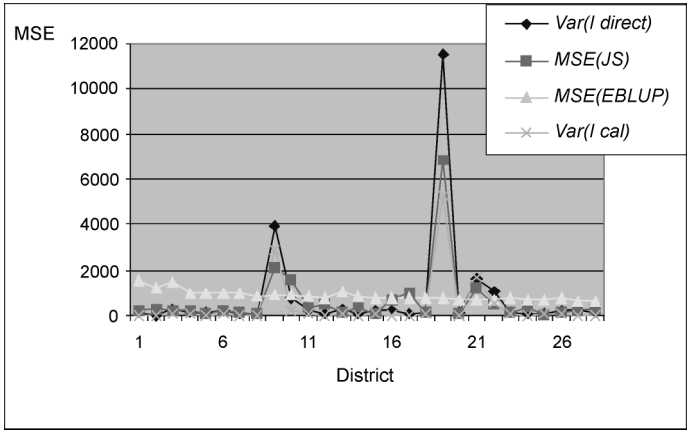

Fig. 2. Accuracy of the estimates of income per capita.

5. Conclusions

Fig. 1 shows that two regression estimates do not differ significantly, as well as two composite estimates.

The averages over small areas of the estimated mean square errors ($\overline{MSE}$) of four estimates of income I are presented in Table 2.

Table 2

Average MSE of estimates

Estimator	*I direct*	*I JS*	*I EBLUP*	*I cal*
$\overline{MSE}$	790	649	877	404

The James–Stein estimator performs better than the direct one.

The MSE of the composite estimator EBLUP performs equally along the areas, improving a very low accuracy of the direct estimator in some areas, however, without any improvement in the average accuracy. It can be explained by the low correlation between the direct estimates in the areas and auxiliary variables. So, some better auxiliary information has to be found.

The currently used calibrated estimator has the smallest $\overline{MSE}$.

References

[1] J. Deville, C.-E. Särndal, O. Sautory, Generalized raking procedures in survey sampling, *JASA*, **88**, 1013–1020 (1993).

[2] R.E. Fay, R.A. Herriot, Estimates of income for small places: an application of James–Stein procedures to census data, *JASA*, **74**, 269–277 (1998).

[3] M. Ghosh M., J.N.K. Rao, Small area estimation: an appraisal, *Statistical Science*, **9**(1), 55–93 (1994).

Mažos srities įverčio pavyzdys imtyje iš baigtinės populiacijos

D. Krapavickaitė

Pateikiami metodai, kuriais, panaudojant papildomą informaciją, galima patikslinti įverčius srityse, kuriose turima maža imtis. Jie panaudojami Lietuvos kaimo gyventojų vidutinių pajamų rajonuose vertinimui.