Estimation strategy for small areas, a case study

Nekrašaitė-Liegė Vilma[1]

[1]Vilnius Gediminas technical university, e-mail: nekrasaite.vilma@gmail.com

Abstract

The purpose of this research is to find optimal strategy (pair of sample design and estimator) for small area estimation. Thus, the definition of a balanced sample and two special cases of balanced samples are presented. The study variable and auxiliary information are time series, thus the different unit level panel-type models are used not only in estimation stage, but and in sample selection stage. The simulation results showed, that the impact of the chosen model is larger for the small domains than for the large ones. Also results showed that the use of the panel type model in sample selection and estimation stages improves the accuracy of the estimate.
Keywords: small area, balanced sample, panel-type model.

1 Introduction

As mentioned by Ghosh & Rao (1994) the term "small area" and "local area" are commonly used to denote a small geographical area, such as a county, a municipality or a census division. They may also describe a "small domain", i.e., a small subpopulation such as a specific age-sex-race group of people within a large geographical area.
The focus on small area estimation (SAE) is made because the demand for data at lower geographic levels is always present, especially from local governments and from businesses needing to make investment, marketing, and location decisions that depend on knowledge of local areas.
The small area problem is usually considered to be treated via estimation (Ghosh & Rao, 1994). However, if the domain indicator variables are available for each unit in the population there are opportunities to be exploited at the survey design stage. Thus in this paper I am interesting in an overall strategy that deals with small area problems, involving both planning sample design and estimation aspects.

2 Main notations

A finite *population* $U = \{u_1, u_2, ..., u_N\}$ of the size N is considered. For simplicity, in the sequel we identify a population element u_k and its index k. Hence $U = \{1, 2, ..., N\}$.
The elements k ($k = 1, \dots, N$) of the population U has two components $y(t)$ and $\mathbf{x}(t)$. The values of these components depends on time. The component $y(t)$ defines the value of a *study variable* (variable of interest) in time t, and the component $\mathbf{x}(t) = \{x_1(t), x_2(t), \dots, x_J(t)\} \in \mathbb{R}^J$ defines the values of the J *auxiliary variables* in time t.
The population is divided into D nonoverlapping *domains* (subpopulations) $U^{(d)}$ of size $N^{(d)}$, where $d = 1, \dots, D$. Domain indicator variables define whether $k \in U$ belongs to a given domain:

$$q_k^{(d)} = \begin{cases} 1, & \text{if } k \in U^{(d)}, \\ 0, & \text{otherwise,} \end{cases} \qquad \forall k \in U, d = 1, \dots, D. \tag{1}$$

The *parameter of interest* is a *domain total* in time moment t:

$$TOT^{(d)}(t) = \sum_{k \in U^{(d)}} y_k(t) = \sum_{k \in U} q_k^{(d)} y_k(t), \quad d = 1, \dots, D; \quad t = 1, 2, \dots \tag{2}$$

To estimate $TOT^{(d)}(t)$, we need information about unknown variable $y(t)$. This information is collected by sampling. The *sampling vector*

$$\underline{\mathbf{S}}(t) = (\underline{S}_1(t), \underline{S}_2(t), \ldots, \underline{S}_N(t)) \tag{3}$$

is a random vector whose elements $\underline{S}_k(t)$ indicate the number selections for k in time t. The distribution of $\underline{\mathbf{S}}(t)$, denoted by $p(.)$, is called a *sample design*. The realization $\mathbf{S}(t) = (S_1(t), S_2(t), \ldots, S_N(t))$ of $\underline{S}_k(t)$ is called *sample*. It define the *sample set* $s(t) = \{k : k \in U, S_k(t) \geq 1\}$.

3 Balanced samples

The sample might be balanced or unbalanced. A sample is said to be balanced if, for a vector of auxiliary variable $\mathbf{z}(t) = \{z_1(t), z_2(t), \ldots, z_L(t)\} \in \mathbb{R}^L$,

$$\sum_{k \in s(t)} \frac{\mathbf{z}_k(t)}{\pi_k(t)} = \sum_{k \in U} \mathbf{z}_k(t). \tag{4}$$

In other words, in a balanced sample, the total of the z-variables are estimated without error. Let us note that two different sets of variables have been introduced in order to underline that the set of variables available at the design stage ($\mathbf{z}$ variables) could be different from the set available at the estimation stage ($\mathbf{x}$ variables) even if in many practical situations they could be the same. Here, an element's k inclusion probability in time t is denoted as $\pi_k(t)$.
Almost all the other sampling techniques are particular cases of balanced sampling. Some well-known sampling designs are particular cases of balanced sampling:
1. Sampling with a fixed sample size is a particular case of balanced sampling. In this case, the only balancing variable is $\pi_k(t)$. The balancing equations given in (4) become

$$\sum_{k \in s} \frac{\pi_k}{\pi_k} = \sum_{k \in s} 1 = \sum_{k \in U} \pi_k,$$

which means that the sample size must be fixed.
2. Stratification is a particular case of balanced sampling. In this case, the balancing variables are the indicator variables of the strata

$$\delta_{kh} = \begin{cases} 1, & \text{if } k \in U_h, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where U_h denotes population part, which belongs to h, $h = 1, ..., H$ strata. Since the inclusion probabilities in stratum h are $\pi_k = n_h/N_h$, $k \in U_h$, the balancing equations become

$$\sum_{k \in s} \frac{N_h \delta_{kh}}{n_h} = \sum_{k \in U} \delta_{kh} = N_h, \quad h = 1, ..., H,$$

and are exactly satisfied.
The main reason why a balanced sample is used in my research is that the Deville & Tillé (2004) showed, that the optimal strategy contains a balanced sample. Thus, to select a balanced sample is not easy. The one of the quickest way to select balanced sample is to use cube method.

3.1 Cube method

The algorithm of the cube method was proposed by Deville & Tillé (1998), and the method was published in Tillé (2001) and Deville & Tillé (2004). This method is general in the sense that the inclusion probabilities are exactly satisfied, that these probabilities may be equal or unequal and that the sample is as balanced as possible.
It is possible to get SAS/IML version of Cube method done by Chauvet & Tillé (2006) and it is also available on the University of Neuchatel Web site. This software program is free, available over the Internet and is easy to use.

4 Model-based sample design

In this paper balanced samples are selected not only using well known simple random sample, stratified simple random sample, but also using model-based sample (Nekrašaitė-Liegė *et al.*, 2011). The suggested model-based sample design consists of three steps:

1. Model construction and estimation of it's coefficients;
2. Estimation of the variance of the prediction error;
3. Construction of the sample design $p(.)$.

In the first step the model is fitted to the available auxiliary data. In the second step the prediction errors (residuals)

$$\hat{\varepsilon}_k(t) = \hat{y}_k(t) - y_k(t), \quad t \in \mathcal{T}_k \in \{1, 2, ..., T.\}, \tag{6}$$

are calculated and the variance of prediction error in each domain, $\sigma^{(d)2}$, is estimated. Finally, in the third step the (approximately) optimal sample design $p(.)$ (actually, the Neyman stratified simple random sample, (Särndal *et al.*, 2003)) based on the estimated variances $\widehat{\sigma^{(d)2}}$ is constructed. Thus, the less model-based prediction accuracy in the domain the more elements from this domain are drawn.

5 Estimators and models

After the sample is selected the domain total is calculated using GREG-type (Lehtonen *et al.*, 2003) estimator:

$$\widehat{TOT}^{(d)}_{GREG}(t) = \sum_{k \in U^{(d)}} \hat{y}_k(t) + \sum_{k \in s(t) \cap U^{(d)}} 1/\pi_k(t)(y_k(t) - \hat{y}_k(t)). \tag{7}$$

where $\hat{y}_k(t)$ denotes the prediction of $y_k(t)$ under the assumed super population model. The predictions $\{\hat{y}_k(k); k \in U\}$ differ from one model specification to another, depending on the functional form and from the choice of the auxiliary variables.
In this paper a general panel data model with random effects is considerate as working super population model:

$$\begin{aligned} Y_k(t) =& \beta_{0,g(k)}(t) + r_{0,k}(t) + \sum_{j=1}^{J} [\beta_{j,g(k)}(t) + r_{j,k}(t)] X_{j,k}(t) + \\ &+ \sum_{i=1}^{m} \alpha_{i,g(k)} \mu_i(t) + \varepsilon_k(t), \quad k \in U. \end{aligned} \tag{8}$$

Here $X_{j,k}(t)$, $j = 1, 2, ..., J$, are fixed-effects variables, $\beta_{0,g(k)}(t)$, $\beta_{1,g(k)}(t)$, ..., $\beta_{J,g(k)}(t)$ are the unknown fixed-effects model coefficients, which are the same in group $g(k)$. The groups $g(k)$ divides population U into G nonoverlaping groups which in some special cases can be the same as domains d, $d = 1, ..., D$. The unknown random-effects models coefficients are denoted as $r_{0,k}(t)$, $r_{1,k}(t)$,..., $r_{J,k}(t)$ ($r_{p,k}(t) \sim IID(0, \lambda^2_{0,g(k)}(t)), g(k) = 1(k), ..., G(k), p = 0, ..., J$. The model error is denoted as $\varepsilon_k(t)$ ($E_M(\varepsilon_k(t)) = 0$, $VAR_M(\varepsilon_k(t)) = \nu_k^2 \sigma^2$, $\forall k \in U$ and $cov(\varepsilon_k(t), \varepsilon_l(t)) = 0$ when $k \neq l$). It should be noticed that model error $\varepsilon_k(t)$ and the random-effects model coefficients $r_{0,k}(t)$, $r_{1,k}(t)$,..., $r_{J,k}(t)$ are conditionally independent if values of $X_{j,k}(t)$, $j = 1, 2, ..., J$, are known. The component $\sum_{i=1}^{m} \alpha_{i,g(k)} \mu_i(t)$ represents a time trend. The structure of this component depends on historical auxiliary information and is specified using exploratory analysis.
Below several special cases of this model are described:

- *Example 1.* Let $\beta_{0,g(k)}(t) = \beta_0(t)$, $r_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $r_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has such form

$$Y_k(W) = \beta_0(W) + \sum_{j=1}^{J} \beta_j(W) X_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \tag{9}$$

This model is known as common model (Lehtonen *et al.* (2003)), because it has the same model parameters for all domains.

- *Example 2.* Let $\beta_{0,g(k)}(t) = \beta_0^{(d)}(t)$, $r_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $r_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has such form

$$Y_k(W) = \beta_0^{(d)}(W) + \sum_{j=1}^{J} \beta_j(W) X_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \tag{10}$$

This model is known as model with domain-intercept (Lehtonen *et al.* (2003)), because it has the same slopes but separate intercepts for all domains.

- *Example 3.* Let $\beta_{0,g(k)}(t) = \beta_{0,g(k)}$, $r_{0,k}(t) = 0$, $\beta_{j,g(k)}(t) = \beta_{j,g(k)}$ $r_{j,k}(t) = 0$, $j = 1, ..., J$. Then the generalized unit level model has such form

$$Y_k(t) = \beta_{0,g(k)} + \sum_{j=1}^{J} \beta_{j,g(k)} X_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \tag{11}$$

This model is fixed effect panel data model. Here models coefficients $\beta_{0,g(k)}$, $\beta_{1,g(k)}$, ..., $\beta_{J,g(k)}$ do not depend on time which means they are the same for the all periods of time. Such model is very useful in practice since it enables one to find model coefficients just using data from the past. The current data might be use just for prediction.

- *Example 4.* Let $\beta_{0,g(k)}(t) = \beta_0(t)$, $r_{0,k}(t) = r_{0,g(k)}(t)$, $\beta_{j,g(k)}(t) = \beta_j(t)$, $r_{j,k}(t) = 0$, $j = 1, ..., J$ and t is equal to one moment (let this moment is notated as W). Then the generalized unit level model has such form

$$Y_k(W) = \beta_0(W) + r_{0,g(k)}(W) + \sum_{j=1}^{J} \beta_j(W) X_{j,k}(W) + \varepsilon_k(W), \quad k \in U. \tag{12}$$

This model is known as mixed model with random-intercept, because it has the same fixed parameters for all domains. The random effect is defined at the group level.

- *Example 5.* Let $\beta_{0,g(k)}(t) = \beta_0^{(d)}$, $r_{0,k}(t) = r_{0,g(k)}$, $\beta_{j,g(k)}(t) = \beta_j^{(d)}$, $r_{j,k}(t) = 0$, $j = 1, ..., J$. Then the generalized unit level model has such form

$$Y_k(t) = \beta_0^{(d)} + r_{0,g(k)} + \sum_{j=1}^{J} \beta_j^{(d)} X_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \tag{13}$$

It is assumed that the model coefficients $\beta_0^{(d)}$, $\beta_j^{(d)}$, $j = 1, ..., J$, and the random effects $r_{0,g(k)}$ do not depend on time (they are the same during the different time periods).

- *Example 6.* Let $\beta_{0,g(k)}(t) = \beta_0^{(d)}$, $r_{0,k}(t) = r_{0,g(k)}$, $\beta_{j,g(k)}(t) = \beta_j^{(d)}$, $r_{j,k}(t) = 0$, $j = 1, ..., J$. Then the generalized unit level model has such form

$$Y_k(t) = \beta_0^{(d)} + r_{0,g(k)} + a_0^{(d)} t + \mathbf{a}'^{(d)} \alpha(t) + \sum_{j=1}^{J} \beta_j^{(d)} x_{j,k}(t) + \varepsilon_k(t), \quad k \in U. \tag{14}$$

This is a panel data model with a linear trend and a seasonal component, $\mathbf{a}^{(d)\prime}\alpha(t)$, $\mathbf{a}^{(d)} \in \mathbf{R}^3$.

6 Simulation and Conclusions

For the simulation experiment, a real population from Statistics Lithuania is used. Enterprisers which are responsible for education are taken as the finite population. Information about these enterprisers is taken 20 times (each quarter from 2005 till 2009). The average number of enterprises in each quarter is 750 (Number of population).
The study variable $y_k(t)$ is the income of an enterprise k and the auxiliary variables are the number of employers $x_{1,k}(t)$, tax of value added (VAT) $x_{2,k}t$ and various indicators (specification of enterprise (5 indicators), size of enterprise (3 indicators), region (6 indicators)) $x_{j,k}$, $j = 3, ..., 15$.
The total income in a domain in each quarter in 2008 and 2009 is chosen as the parameter of interest ($T + l$, $T = 12$, $l = 1, ..., 8$). The domain is chosen as counties (there are 10 counties in Lithuania) and specification of enterpriser (5 specifications). So, in this research the study variables are elements of a time series with 8 elements and the total number of domains of interest is 120. The number of enterprises in each domain varies from 6 to over than 300.
The comparison of results of 1000 simulations using different models and auxiliary information in both stages (sample selection and estimation) showed that the impact of the model is larger for the small domains than for the large domains. Also results showed that the use of the panel type model in sample selection and estimation stages improves the accuracy of the estimate. More results and conclusions will be presented during presentation.

References

Chauvet, G. & Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics* **21**, 9 – 31.

Deville, J.-C. & Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89 – 101.

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893 – 912.

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

Lehtonen, R., Särndal, C.-E. & Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology* **29**, 33 – 44.

Nekrašaitė-Liegė, V., Radavičius, M. & Rudys, T. (2011). Model-based design in small area estimation. *Lithuanian mathematical journal* **51**, 417 – 424.

Särndal, C., Swensson, B. & Wretman, J. (2003). *Model assisted survey sampling.* Springer Verlag.

Tillé, Y. (2001). *Thorie des sondages : chantillonnage et estimation en populations finies.* Dunod, Paris.