STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 265—280

SOME APPLICATIONS OF PANEL DATA MODELS IN SMALL AREA ESTIMATION

Vilma Nekrašaitė-Liegė[1]

ABSTRACT

This study uses a real population from Statistics Lithuania to investigate the performance of different types of estimation strategies. The estimation strategy is a combination of sampling design and estimation design. The sampling designs include equal probability design (SRS) and unequal probability designs (stratified SRS and model-based sampling designs). Design-based direct Horvitz-Thompson, indirect model-assisted GREG estimator and indirect model-based estimator are used to estimate the totals in small area estimation. The underlying panel-type models (linear fixed-effects type or linear random-effects type) are examined in both stages of estimation strategies: sample design and construction of estimators.

Key words: Small area estimation; panel-type data; model-based; model-assisted

1. Introduction

In this paper, the accuracy of several small area estimation strategies (a pair comprising a sample design and an estimator) is investigated. The focus on small area estimation (SAE) is made because SAE is an important objective of many surveys. Small areas almost always have small sample sizes, so standard survey estimation methods, which only use information from the small area samples, are unreliable for these areas. In this context SAE methods that borrow strength via statistical models (Rao 2003) are used to produce reliable estimates.

Nowadays, official statistics repeats the same surveys from year to year, so for most of the population elements it is possible to get information for the same variable in several time periods. It means that for many surveys individual data for some objects are known for at least one previous time point. We will call this panel-type data even when it is not part of the design. Also, in some cases it is possible to use information collected from the other sources (tax offices,

[1] Vilnius Gediminas Technical University, Vilnius. E-mail: nekrasaite.vilma@gmail.com

jobcentres, etc.). Such dataset of a large amount of auxiliary information might improve the quality of the estimation strategy as compared with a strategy based on the current sample alone.

The use of panel-type data in estimation strategy means that a prediction theory based on a superpopulation model is used. A superpopulation model can be used not only in estimation stage, but for sample selection as well. Such use of the superpopulation model is discussed for example by Royall (1970) and Nedyalkova and Tille (2008). For the estimation strategy they used linear regression model as a superpopulation model. In our research, a basic superpopulation model is an incomplete panel data model (Hsiao 2003).

The advantage of using a panel data model in model-based sample design for small areas has been noticed by Nekrašaitė-Liegė, Radavičius and Rudys (2011). In this research we place emphasis on application of high-dimensional multi-level fixed effect panel data model using comprehensive exploratory analysis and model selection technique. The results obtained using such model are compared with the results obtained using panel data model with random effect for domains and panel data model with few fixed effect that are the same for the whole population.

In this research not only different panel data models are compared, but also two types of estimators: model-assisted and model-based. Almost all papers devoted to model-based estimation in small areas (see, e.g., Rao 2005, Omrani, Gerber and Bousch 2009) deal with samples from a limited number of areas and the case where a set of auxiliary covariates are used to obtain estimates in the areas that are not actually sampled. In this research a sample from all areas is selected (number of selected elements in the areas is random) and a set of auxiliary information is known for all elements in the population from the previous surveys and other administrative sources. Such type of auxiliary information might be available in short term statistics where, at a current time, data are collected using small samples and in the future the data for the same period of time is expanded and updated using large samples or administrative sources. Thus the use of such type of auxiliary information might help us to find elaborate fixed effect panel data model appropriate for the effective model-based estimation strategy.

Hence, in this paper several estimation strategies are used to answer the following problems: what type of model (with fixed or random effects), sample design and estimator (design-based or model-based) should be used in small area estimation.

The paper consists of six sections. The main notation and definitions used in survey statistics are introduced in Section 2. In Section 3, different types of estimators and estimates are presented. Some panel data models applicable in survey sampling are described in Section 4. We end with simulation results (Section 5) and concluding remarks (Section 6).

2. Definitions and notations

Let us start with a common framework of finite population survey sampling. 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 as its index k . Hence $U=\{1, 2, ..., N\}$.

The elements k (k=1, 2, ..., N) of the population U has two components y and $\mathbf{x}$.

The component y defines the value of a study variable (variable of interest), and the component $\mathbf{x}=\{x_1, x_2, ..., x_J\}\in\boldsymbol{R}^J$ defines the values of the J auxiliary variables.

In this study a panel-type data is considered. This means that y_k and $\mathbf{x}=\{x_{1,k}, x_{2,k}, ..., x_{J,k}\}$ are assumed to be time series,

$$y_k=(y_k(t), t=1,2,...),\quad \mathbf{x}_k=(\mathbf{x}_k(t), t=1,2,...). \tag{1}$$

The population is divided into D nonoverlapping domains (subpopulations) U(d) of size N(d), where d=1, 2, ..., D. Domain indicator variables define whether k ∈ U belongs to a given domain:

$$q_k^{(d)}=\begin{cases}1, & \text{if } k\in U^{(d)}\\ 0, & \text{otherwise}\end{cases}, \forall k\in U, d=1,...,D. \tag{2}$$

It is assumed, that an element can not change domain during the time period, thus $q_k^{(d)}$ do not depend on time.

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

$$T^{(d)}(t)=\sum_{k\in U} q_k^{(d)} y_k(t)=\sum_{k\in U^{(d)}} y_k(t),\quad d=1,...,D. \tag{3}$$

Actually in this study, the parameter of interest is a domain total in 4 different quarters for a given year. If time variable t denotes quarters and T+1 denotes the first quarter of interest, then the parameter of interest is

$$T^{(d)}(T+l)=\sum_{k\in U} q_k^{(d)} y_k(T+l)=\sum_{k\in U^{(d)}} y_k(T+l),\quad d=1,...,D,\quad l=1,...,4. \tag{4}$$

To estimate $T^{(d)}(T+l)$, we need information about unknown variable y in time $T+l$. This information is collected by sampling. The sampling vector $\underline{\mathbf{S}}(T+l)$ = $(\underline{S}_1(T+l), \underline{S}_2(T+l), ..., \underline{S}_N(T+l))$ is a random vector whose elements $\underline{S}_k(T+l)$ indicate the number of selections for k element in time point $T+l$. In this research we are interested just in the sampling without replacement (WOR), thus the largest number of selections for k element in $T+l$ is one: $\underline{S}_k(T+l)=1$ if element k is selected and $\underline{S}_k(T+l)=0$ if it is not selected. Also, in this paper, the sampling vector is the same for all 4 quarters of interest, thus it can be notated as $\underline{\mathbf{S}}(T+l)=\underline{\mathbf{S}}=(\underline{S}_1, \underline{S}_2, ..., \underline{S}_N)$, $l=1,\ldots,4$. It means that the sampling vector is determined for the first quarter of interest and it is repeated for other 3 quarters. The realization $\mathbf{S}(T+l) = \mathbf{S} = (S_1, S_2, ..., S_N)$ is called a sample. Let $\boldsymbol{S}$ be the set of all samples $\mathbf{S}$. The sampling vector $\underline{\mathbf{S}}$ (and its realization $\mathbf{S}$) define the sample set $\underline{s}$ (and the corresponding s) as

$$\underline{s}(T+l)=\underline{s}=\{k:k\in U,\underline{S}_k=1\},\quad s(T+l)=s=\{k:k\in U,S_k=1\}. \tag{5}$$

The difference between sample $\mathbf{S}$ and sample set s is that s is a subset of U whereas $\mathbf{S}$ is a N-dimensional vector of indicators.

The distribution of $\underline{\mathbf{S}}$, denoted by $p(\cdot)$, is called a sample design. The sampling design assigns a probability $\mathbf{P}(\underline{\mathbf{S}} = \mathbf{S}) = p(\mathbf{S})$ to every sample $\mathbf{S}$. First and second order inclusion probabilities π_k and π_{kl} for sampling without replacement (WOR) are defined as

$$\pi_k=\mathbf{P}(\underline{S}_k=1)=\sum_{\mathbf{S}:S_k=1}p(\mathbf{S})=\mathbf{E}(\underline{S}_k),$$

(6)

$$\pi_{kl}=\mathbf{P}(\underline{S}_k=1,\underline{S}_l=1)=\sum_{\mathbf{S}:S_kS_l=1}p(\mathbf{S})=\mathbf{E}(\underline{S}_k,\underline{S}_l), \tag{7}$$

where $\pi_{kk} = \pi_k$ and $\operatorname{cov}(\underline{S}_k,\underline{S}_l)=\Delta_{kl}=\pi_{kl}-\pi_k\pi_l$. Thus the samples for each quarter are the same as for the first quarter, the inclusion probabilities do not depend on time.

The sampling weights for WOR designs are defined as

$$w_k=\begin{cases}\pi_k^{-1}, & \text{if } k\in s\\ 0, & \text{otherwise}\end{cases}. \tag{8}$$

The sample size and the sample set in domain $U^{(d)}$ are

$$n^{(d)}=\sum_{k\in U^{(d)}}S_k,\quad s^{(d)}=s\cap U^{(d)}. \tag{9}$$

There are two types of domains:

1. Planned domains. (Singh, Gambino and Mantel 1994) For planned domains the sample size $n^{(d)}$ in domain sample is fixed in advance, so really these domains are strata with possible different allocations.

2. Unplanned domains. If the sample size $n^{(d)}$ in domain sample is random, domains are unplanned. The disadvantage of unplanned domains is that, there might be domains with zero elements in the sample **S**.

In this research domains are unplanned. It is assumed that the number of the elements in each domain $U^{(d)}$, d=1, 2, ..., D, is known, but the domains are not used in the sample design. This means that the sample part in each domain, $s^{(d)}$, has a random size.

3. Estimators and estimates

An estimator is a rule or algorithm that defines how to estimate the parameter of interest (in our case: domain total). It is a random variable, which value depends on the sample and the auxiliary information. An estimate is the realized value of an estimator. In general, an estimator and an estimate are denoted, respectively, as $\hat{\theta}(\underline{\mathbf{S}})$ and $\hat{\theta}(\mathbf{S})$, or briefly as $\underline{\hat{\theta}}$ and $\hat{\theta}$. For parameter $T^{(d)}(T+l)$, the estimator and estimate are $\underline{\hat{T}}^{(d)}(T+l)$ and $\hat{T}^{(d)}(T+l)$, l=1,…,4.

The estimator is accurate if its bias and variance are small. The bias is the difference between the parameter expectation and the true value: $BIAS(\underline{\hat{\theta}}) = \mathbf{E}(\underline{\hat{\theta}}) - \theta$. If $BIAS(\underline{\hat{\theta}}) = 0$, the estimator is unbiased. The bias might come with respect to the design or to the model. The symbols **E**, *var* denote, respectively, the expected value and the variance under the sample design. They are defined as

$$\mathbf{E}(\underline{\hat{\theta}}) = \sum_{\boldsymbol{S} \in \boldsymbol{S}} p(\mathbf{S})\,\hat{\theta} \tag{10}$$

$$var(\underline{\hat{\theta}}) = \sum_{\boldsymbol{S} \in \boldsymbol{S}} p(\mathbf{S})\,[\hat{\theta} - \mathbf{E}(\underline{\hat{\theta}})]^2 \tag{11}$$

In this research two types of estimators of the domain total are used:

1. Design-based estimators. The design-based estimators can be divided in two groups (Särndal, Swensson and Wretman 1992, Lehtonen and Veijanen 2009): design-based direct estimators, which are design unbiased by definition and design-based model-assisted indirect estimators, which are nearly design unbiased irrespective of the model choice.

2. Model-based estimators. A model-based estimator usually has smaller variance then a design-based estimator, and it is possible to use them even when there is no selected unit in the domain. Still model-based estimator is design- biased and in some cases it might have a large bias.

Two types of estimators can be used for estimation in domains:

1. Direct estimators. A direct estimator uses values of the variable of interest only from the time period of interest and only from units in the domain of interest (U.S. office of management and budget 1993).
2. Indirect estimators. An indirect domain estimator uses values of the variable of interest from a domain and/or time period other than the domain and time period of interest (U.S. office of management and budget 1993).

A convenient direct estimator is Horvitz - Thompson (HT) estimator (Narain, 1951, and Horvitz and Thompson, 1952) for the domain $\hat{\underline{T}}^{(d)}_{HT}(T+l)=\sum_{k\in\underline{s}^{(d)}} w_k y_k(T+l)$ and it‘s estimate is $\hat{T}^{(d)}_{HT}(T+l)=\sum_{k\in s^{(d)}} w_k y_k(T+l)$, l=1,…,4.

Another estimator is the generalized regression (GREG) estimator (Särndal, Swensson and Wretman 1992). The estimator and estimate for the domain total are

$$\hat{\underline{T}}^{(d)}_{GREG}(T+l)=\sum_{k\in\underline{U}^{(d)}}\hat{y}_k(T+l)+\sum_{k\in\underline{s}^{(d)}} w_k(y_k(T+l)-\hat{y}_k(T+l)) \text{ and}$$
$$\hat{T}^{(d)}_{GREG}(T+l)=\sum_{k\in U^{(d)}}\hat{y}_k(T+l)+\sum_{k\in s^{(d)}} w_k(y_k(T+l)-\hat{y}_k(T+l)) \qquad (12)$$

The last estimator is Model-based (MB) estimator defined by

$$\hat{\underline{T}}^{(d)}_{MB}(T+l)=\sum_{k\in\underline{U}^{(d)}\setminus\underline{s}^{(d)}}\hat{y}_k(T+l)+\sum_{k\in\underline{s}^{(d)}} y_k(T+l) \text{ and}$$
$$\hat{T}^{(d)}_{MB}(T+l)=\sum_{k\in U^{(d)}\setminus s^{(d)}}\hat{y}_k(T+l)+\sum_{k\in s^{(d)}} y_k(T+l). \qquad (13)$$

For the both (GREG and MB) estimators, $\hat{y}_k(T+l), k\in U^{(d)}$, are predicted values of study variable y for each element in $U^{(d)}$ in time $T+l$, l=1,…,4. The prediction algorithm is described in Section 4. Due to the prediction algorithm GREG and MB estimators are indirect. Thus in this paper direct design-based (HT), indirect design-based (GREG) and indirect model-based (MB) estimators are compared.

4. Panel data models in survey sampling

The use of model-assisted or model-based estimators is impossible unless some model is considered. In the most papers a linear regression model is exploited (see, e.g., Royall 1970, Nedyalkova, Tille 2008, and references therein), however in some cases a generalized linear mixed model (Saei and Chambers

2003, Lehtonen, Särndal and Veijanen 2003, 2005, Lehtonen and Veijanen 2009) is applied.

In this study a panel-type data is considered. The problem is to find an effective strategy for estimating the totals of $y_k(T+l)$, $k \in U^{(d)}$, l=1,…,4, given the ("historical", i.e. prior to the sample selection) auxiliary information

$$AI := \left(\mathbf{x}_k(t), y_k(t), t \in \mathrm{T}_k \subset \{1,2,...,T\}, k \in U\right). \tag{14}$$

Let $y_k(t)$ and $\mathbf{x}_k(t)$, k = 1, ..., N, be the realizations of random variables $\underline{y}_k(t)$ and $\underline{\mathbf{x}}(t)=\{\underline{x}_{1,k}(t), \underline{x}_{2,k}(t), ..., \underline{x}_{J,k}(t)\}$ of the superpopulation model $\boldsymbol{M}$:

$$y_k(t) = \beta_{0,g(k)}(t) + r_{0,k}(t) + \sum_{j=1}^{J} \beta_{j,g(k)}(t)\, \underline{x}_{j,k}(t) + \sum_{i=1}^{m} \alpha_{i,g(k)}\mu_i(t) + \varepsilon_k(t), \;\; k \in U. \tag{15}$$

Here $\underline{x}_{j,k}(t)$, 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. The unknown random-effects models coefficient is denoted as $r_{0,k}(t)$ $\left(r_{0,k}(t) \sim IID\left(0, \lambda^2_{0,g(k)}(t)\right), g(k) = 1,..., G(k)\right).$

The model error is denoted as $\varepsilon_k(t)$ $\mathbf{E}_\mathbf{M}\left(\varepsilon_k(t)\right) = 0, var_\mathbf{M}\left(\varepsilon_k(t)\right) = \upsilon_k^2\sigma^2, \forall k \in U$ and $\mathrm{cov}\left(\varepsilon_k(t), \varepsilon_l(v)\right) = 0$ when $(k,t) \neq (l,v)$). It should be noticed that model error $\varepsilon_k(t)$ and the random-effects model coefficient $r_{0,k}(t)$ are conditionally independent if values of $\underline{x}_{j,k}(t)$, j=1, 2, ..., J, are given. 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 (14) and is specified using exploratory analysis.

Some special cases of general panel data model (15) are the following:

1. Fixed effect panel data model:

$$\underline{y}_k(t) = \beta_{0,g(k)} + \sum_{j=1}^{J} \beta_{j,g(k)}\, \underline{x}_{j,k}(t) + \sum_{i=1}^{m} \alpha_{i,g(k)}\mu_i(t) + \varepsilon_k(t), \;\; k \in U. \tag{16}$$

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.

2. Random effect panel data model:

$$\underline{y}_k(t) = \beta_{0,g(k)} + r_{0,k} + \sum_{j=1}^{J} \beta_{j,g(k)}\, \underline{x}_{j,k}(t) + \sum_{i=1}^{m} \alpha_{i,g(k)}\mu_i(t) + \varepsilon_k(t), \;\; k \in U. \qquad (17)$$

Here the random effect $r_{0,k}$ is included into the previous model. Random effect also does not depend on time and hence it is also possible to find all model coefficients from the past data.

Thus the use of models which coefficients are known before the sample is selected might improve not only the estimators but the sample design as well. The application of the same model in both stages (sample selection and estimation) might be very useful.

5. Simulation

5.1. Population

For the simulation experiment, a real population from Statistics Lithuania is used. Enterprisers which are responsible for adult and other education and have less than 50 employers are taken as the finite population. Information about these enterprisers is taken 16 times – each quarter from 2005 till 2008. The average number of enterprises in each quarter is 650 (Number in population).

The study variable y_k is the income of an enterprise k and the auxiliary variables are the number of employers $x_{1,k}$, tax of value added (VAT) $x_{2,k}$ and various indicators (specification of enterprise (5 indicators), size of enterprise (2 indicators)) $x_{\mathrm{j},k}$, $j = 3, \ldots, 9$.

The study parameter is the total income $T^{(d)}$, in the domain d. The domain is chosen as counties (there are 10 counties in Lithuania). The number of enterprises in each domain varies from 6 to 323 (see table 1.).

Table 1. Domain size in population

Domain size	Number of enterprisers in domain	Number of domains in one quarter	Total number of domains of interest
Small	6 – 25	5	20
Medium	25 – 50	2	8
Large	>50	3	12

The total income in a domain in each quarter in 2008 is chosen as the parameter of interest ($T + l$, T=12, l=1,...,4). So, in this research the study variables are elements of a time series with 4 elements and the total number of domains of interest is 40 (see table 1.).

The overall available auxiliary information is divided into two sets: the "historical" data *AI* (formula (14)) available before the sample selection, i.e. in the sample design stage, and new auxiliary information with the true observations

$$AI(l) := \begin{pmatrix} \mathbf{x}_u(T+l), u \in U \\ y_k(T+l), k \in s \end{pmatrix}, \tag{18}$$

which is available at estimation stage for each quarter *l* under consideration (*l*=1,...,4).

5.2. Estimation strategy

Before selecting a sample, the three different panel-type data models were analyzed using *AI*. A detailed exploratory data analysis has been performed in order to construct an appropriate model for the data. For instance, model (FI) has been selected from a quite large set of alternative models using model selection technique. In particular, panel models with the enterprise-specific slopes and/or the seasonal components have been tried out. These models are as follows:

1. Linear fixed effect panel data model (FC):

$$\underline{y}_k(t) = \beta_{0,h} + \sum_{j=1}^{3} \beta_{j,h}\, \underline{x}_{j,k}(t) + \sum_{i=1}^{3} \alpha_{i,h} s_i(t) + \varepsilon_k(t), \;\; k \in U. \tag{19}$$

Here, the index $h \in \{1,2\}$ denotes the size of an enterprise (small or medium), auxiliary variables are the number of employers $\underline{x}_{1,k}(t)$, tax of value added $\underline{x}_{2,k}(t)$ and $\underline{x}_{3,k}(t)$, which indicates whether the enterprise engages in a specific activity (learning to drive) or not. The variable s_i, $i \in \{1,2,3\}$, is the indicator of the i-th quarter.

2. Linear mixed panel data model with domain-specific random effects (RD):

$$\underline{y}_k(t) = \beta_{0,h} + r_o^{(d)} + \sum_{j=1}^{3} \beta_{j,h}\, \underline{x}_{j,k}(t) + \sum_{i=1}^{3} \alpha_{i,h} s_i(t) + \varepsilon_k(t), \;\; k \in U. \tag{20}$$

The difference between RD and FC model is the additionally included random effect $r_0^{(d)}$, $\left(r_0^{(d)} \sim IID\left(0, \lambda_0^{(d)2}\right)\right)$ for domains.

3. Fixed effect panel data model with different intercepts for enterprisers (FI):

$$\underline{y}_k(t) = \beta_{0,k} + \sum_{j=1}^{2} \beta_{j,h}\, \underline{x}_{j,k}(t) + \gamma_{1,h} s_1(t) \underline{x}_{2,k}(t) + \sum_{j=3}^{8} \gamma_{2,1}^{(d)} s_1(t)\, \underline{x}_{j,k}(t) + \sum_{i=1}^{3} \alpha_{i,h} s_i(t) + \varepsilon_k(t),\ \ k \in U. \qquad (21)$$

Here the intercept $\beta_{0,k}$ is different for each enterprise, the component $\gamma_{1,h} s_1(t) \underline{x}_{2,k}(t)$ indicates the difference of $\underline{x}_{2,k}(t)$ in the first quarter and the component $\sum_{j=3}^{8} \gamma_{2,1}^{(d)} s_1(t) \underline{x}_{j,k}(t)$ represents the difference between small enterprise specifications in the first quarter. This effect was revealed in the explorative analysis of "historical" data.

Using these three models a model-based sample design is applied (Nekrašaitė-Liegė, Radavičius, Rudys 2011). It consists of three steps.

1. In the first step (FC) model is fitted to the available auxiliary information *AI*.
2. In the second step the prediction errors (residuals) $\varepsilon_k(t) = \hat{y}_k(t) - y_k(t), t \in \mathrm{T}_k$ are calculated and the variance of prediction error $var_{\mathbf{M}}(\varepsilon_k(t)), \forall k \in U$ for each enterpriser is estimated. This is possible to do, because $var_{\mathbf{M}}(\varepsilon_k(t))$ does not depend on time (see formula (15)).
3. Finally, in the third step the (approximately) optimal sample design $p(\mathbf{S})$ based on the estimated variances is constructed. In this case, the stratified probability proportional to size variable sample design is used where the size variable is the variance of prediction error for each enterpriser. Thus the less model-based prediction accuracy for the enterprise the greater its probability to be selected into the sample.

The same is done and using (RD) and (FI) models. Hence three different model-based (MB-FC, MB-RD, MB-FI) sample designs are used for selecting sample. A sample of n=230 enterprisers is selected for the whole 2008 year, thus the selected enterprisers are the same for all 4 quarters. Since the performance of estimation is investigated for one year, the rotation has no effect and is not considered in the paper.

For the comparison, two more sample designs are constructed: Simple random sampling with n=230 enterprisers and Stratified simple random sampling with the same number of enterprisers. For stratification the size of enterprisers is used to define strata. There are two strata: small (160 enterprisers are selected from 545) and medium (70 enterprisers are selected from 105. All sample designs are without replacement, i.e. each enterprise cannot be selected more than one time in one sample.

For each sample design, three types of estimators are considered: Horvitz - Thompson (HT), Generalized regression (GREG, see equation (12)) and Model-based (MB, see equation (13)) estimators. For the last two estimators, the predicted values are calculated in three ways using (FC), (RD) and (FI) models, respectively. The model coefficients are estimated using the auxiliary information *AI*. Thus, model coefficients are the same for all quarters, the auxiliary information with true observations *AI*(*l*) is used just for estimation of the predicted values.

5.3. Simulation results

To compare the performance of the different estimators (the estimation strategies) a design-based relative root mean squared error (*RRMSE*) for *M* = 1000 simulations is evaluated:

$$RRMSE(t) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{T}_m^{(d)}(t) - T^{(d)}(t)\right)^2}}{T^{(d)}(t)}. \qquad (22)$$

Here $\hat{T}_m^{(d)}(t)$ is the estimate of the total for *m*-th simulation in the domain *d* and $T^{(d)}(t)$ refers to the true population total in the same domain. There are 40 domains of interest, so for the better comparison these regions are grouped into three domain sample size classes by the average number of elements in the domain sample (small 0 – 9, medium 10 – 39 and large >40). A mean of relative root means square error (*MRRMSE*) in each class is calculated (see Tables 2–4).

Table 2. HT estimator results

Estimator	Sample design	*MRRMSE* in domains		
		Small domains	Medium domains	Large domains
HT	MB-FI	36,6	21,7	12,6
	MB-FD	36,8	21,8	12,7
	MB-RD	36,7	21,9	12,6
	SRSS	37,9	23,8	15,9
	SRS	40,4	29,1	20,7

Table 3. GREG estimator results

Estimator, model	Sample design	*MRRMSE* in domains		
		Small domains	Medium domains	Large domains
GREG, FI model	MB-FI	14,8	10,3	6,7
	SRSS	15,0	11,8	7,7
	SRS	16,0	12,5	9,5
GREG, FC model	MB-FC	16,8	13,6	8,3
	SRSS	15,7	12,5	8,8
	SRS	15,9	14,4	11,4
GREG, RD model	MB-RD	18,7	13,5	6,5
	SRSS	15,4	12,4	8,7
	SRS	15,8	14,1	11,2

Table 4. MB estimator results

Estimator, model	Sample design	*MRRMSE* in domains		
		Small domains	Medium domains	Large domains
MB, FI model	MB-FI	11,0	11,4	15,1
	SRSS	11,4	9,9	14,9
	SRS	12,2	10,3	14,4
MB, FC model	MB-FC	21,7	25,6	24,5
	SRSS	21,3	23,4	22,6
	SRS	21,4	23,9	21,7
MB, RD model	MB-RD	21,4	26,3	24,4
	SRSS	21,3	24,9	22,7
	SRS	21,0	24,8	21,6

The results for the HT estimators show that the best sample design strategy is model-based strategy. Nevertheless, the accuracy of the HT estimator is twice less than the GREG estimator. For the other two estimators, it is difficult to indicate the best strategy using *RRMSE*. It seems, however, that overall performance of GREG estimator is better whereas MB estimator is better only for the FI model in case of the small domains.

The performance of the hypothesis testing of equality of two variances is taken as an additional criterion for the comparison. Sample designs under the different models and models under the different sample designs are compared (see tables 5–6).

Table 5. Sample designs comparison

Sample design	GREG estimator			MB estimator		
	domains where *var* is significant smaller, %			domains where *var* is significant smaller, %		
	Small domains	Medium domains	Large domains	Small domains	Medium domains	Large domains
FI model						
MB-FI vs SRSS	35,0	50,0	50,0	80,0	100,0	100,0
MB-FI vs SRS	40,0	62,5	75,0	80,0	100,0	100,0
SRSS vs SRS	35,0	50,0	100,0	80,0	50,0	100,0
FC model						
MB-FC vs SRSS	40,0	12,5	58,3	40,0	50,0	100,0
MB-FC vs SRS	40,0	37,5	66,7	40,0	50,0	100,0
SRSS vs SRS	35,0	50,0	100,0	40,0	50,0	100,0
RD model						
MB-RD vs SRSS	20,0	25,0	75,0	40,0	50,0	100,0
MB-RD vs SRS	15,0	50,0	91,7	40,0	50,0	100,0
SRSS vs SRS	40,0	62,5	100,0	60,0	50,0	100,0

Table 6. Models comparison

Models	GREG estimator			MB estimator		
	domains where *var* is significant smaller, %			domains where *var* is significant smaller, %		
	Small domains	Medium domains	Large domains	Small domains	Medium domains	Large domains
MB sample design						
FI vs FC	55,0	37,5	25,0	80,0	50,0	66,7
FI vs RD	40,0	37,5	50,0	60,0	50,0	66,7
RD vs FC	60,0	25,0	8,3	100,0	50,0	33,3
SRSS sample design						
FI vs FC	50,0	37,5	8,3	100,0	100,0	66,7
FI vs RD	50,0	37,5	8,3	100,0	100,0	66,7
FC vs RD	30,0	12,5	0,0	40,0	0,0	0,0
SRS sample design						
FI vs FC	45,0	50,0	16,7	100,0	100,0	66,7
FI vs RD	55,0	37,5	16,7	100,0	100,0	33,3
FC vs RD	30,0	12,5	0,0	40,0	0,0	0,0

Tables 5 and 6 show a percentage of domains, for each model and each sample design, respectively, with significantly smaller variance of the estimators in the first case (respectively, the sample design or the model) as compared to the second. For the rest of domains, the hypothesis about equality of the two variances is not rejected.

The comparison of the sample designs demonstrates that for the GREG estimator and FI model, the use of MB-FI sample design reduces the variance of the estimator for 40% of the small domains and 50% of the medium or the large domains as compared with SRSS or SRS designs. The variance of the MB estimator is smaller for more domains than the GREG estimator for the same model and sample design. For the large domains, using SRSS instead of SRS always reduces the variance for both (GREG and MB) estimators, however, for the small domains the efficiency of SRSS design is much smaller (in some cases it is just 40%).

The comparison of the models shows that the impact of the model is larger for the small domains (especially for the MB estimator) than for the large domains.

The results in table 6 demonstrate that the best prediction model is FI for both estimators.

The analysis of tables 5–6 reveals that the best strategy is to use MB-FI sample design with FI model for both (GREG and MB) estimators and the analysis of tables 3–4 shows that for the small domains the *RRMSE* of MB estimator is quite small, however for the large domains GREG estimator is the best.

6. Concluding remarks

In this paper three different models have been used. The fixed effect panel data model for the whole population (FC) has been taken as the basic panel data model. Then the model has been extended by adding a random effect for domains (RD). This extension has affected the variance of the small domains – it is smaller for more than 30% of the domains (Table 6.). For the large domains, the impact is negligible. A detailed exploratory analysis has been performed in order to improve the underlying panel data model. An attempt to identify enterprise-dependent and significant fixed effects has been made resulting in a smaller variance for more than 50% of the domains (especially when the model-based estimator is used). Nevertheless, for some domains the differences between the models do not significantly affect the variance of the estimators. A reasonable explanation of this observation is that there were structural changes in some enterprisers in 2008 and these changes are not captured by the fitted model. Thus constructing a design-based (nonparametric) test for structural changes in the population is a challenging problem for the further research.

Another aspect investigated in this research, is choice of sample design. The results (Table 5) show that for more than half domains (especially for large ones) the model-based sample design decrease the variance of all estimators (including HT (Table 2)) in particular when FI model is fitted.

This investigation confirms, for the case of panel data model, an empirical observation (Lehtonen, Särndal and Veijanen 2003, 2005) that the design-based estimators (especially model-assisted) show in practice a better design-based performance than model-based estimators. The model-assisted approach enables one to avoid a large bias of an estimator even when there are only few selected elements in the small area. When there are no selected elements in a small area – a model-based estimator is the only choice.

In summary, the comparison of different estimation strategies for the real Lithuanian data has shown that, in the case where a large amount of panel type data is available, the estimation strategy with the FI model based design and the model-assisted estimator (GREG) might be a reasonable choice in small area estimation.

REFERENCES

HORVITZ, D. G. AND THOMPSON, D. J., (1952). A generalization of sampling without eplacement from a finite universe, *Journal of the American Statistical Association*, 47:663–685.

HSIAO, C. (2003). Analysis of Panel Data, Economic Society monographs no.

34, 2nd edition, New York: Cambridge University Press.

LEHTONEN, R. AND VEIJANEN, A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in C.R. Rao and D. Pfeffermann (Eds.) *Handbook of Statistics*, Vol 29B, *Sample Surveys: Inference and Analysis.* New York: Elsevier.

LEHTONEN, R., SÄRNDAL, C.-E. AND VEIJANEN, A. (2003). The effect of model choice in estimation for domains, including small domains, *Survey Methodology* 29:33-44.

LEHTONEN, R., SÄRNDAL, C.-E. AND VEIJANEN, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains, *Statistics in Transition* 7:649-673.

NARAIN, R. D., (1951). On sampling without replacement with varying probabilities, *Journal of the Indian Society of AgriculturalStatistics*, 3:169–174.

NEDYALKOVA, D., TILLE, Y., (2008). Optimal sampling and estimation strategies under the linear model, *Biometrika* 95[3]:521-537.

NEKRAŠAITĖ-LIEGĖ, V., RADAVIČIUS, M. and RUDYS, T. (2011). Model-based design in small area estimation. *Lithuanian Mathematical Journal.* 51[3]:417-424.

OMRANI, H., GERBER, P. AND BOUSCH. P (2009). Model-Based Small Area Estimation with application to unemployment estimates, *World Academy of Science, Engineering and Technology* 49, 793-800

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

Rao, J.N.K. (2005). Inferential issues in small area estimation: some new developments. *Statistics in Transition*, **7**, 513-526.

ROYALL, R. M., (1970). On finite population sampling theory under certain linear regression models *Biometrika* 57[2]:377-387.

SAEI, A. and CHAMBERS, R. (2003). Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects. Methodology Working Paper No. M03/15. University of Southampton, UK.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*, Springer - Verlag, New York.

SINGH, M.P., GAMBINO, J. and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 314.

U.S. OFFICE OF MANAGEMENT AND BUDGET (1993). Indirect Estimators in Federal Programs, *Statistical Policy Worlang Paper 21,* NATIONAL Technical Information Service, Springfield, Virginia.