DOI 10.1007/s10986-015-9277-9
Lithuanian Mathematical Journal, Vol. 55, No. 2, April, 2015, pp. 243–254

Small area estimates for the fraction of the unemployed

Danutė Krapavickaitė [a] **and Tomas Rudys** [b]

[a] Vilnius Gediminas Technical University, Saulėtekio ave. 11, SRL-I, 425, LT-10223 Vilnius, Lithuania
[b] Vilnius University, Akademijos str. 4, LT-08663 Vilnius, Lithuania
(e-mail: danute.krapavickaite@vgtu.lt; tomas.rudys@mii.vu.lt)

Received September 1, 2014; revised March 5, 2015

Abstract. One of the main research trends in contemporary survey sampling and the need to improve the accuracy of the Lithuanian Labor force survey estimates in small geographic areas have stimulated this study. The aim of the paper is to compare area level models and estimation methods for the fraction of the unemployed using simulation based on the Lithuanian Labor Force Survey data. The Fay–Herriot area level model, estimated by empirical best linear unbiased prediction, and the unmatched logit-normal-normal and binomial-logit-normal models, estimated using hierarchical Bayes analysis, are applied. Bayesian imputation is used for areas without sample data. We suggest the composition of some model elements.

MSC: primary 62D05; secondary 62F15

Keywords: finite population, small area estimation, best linear unbiased prediction, hierarchical Bayes analysis

1 Introduction

Estimation of the finite-population parameters in areas with small sample size is one of the hottest current topics in survey sampling. It is not enough to estimate parameters for the whole finite population in real surveys; estimates for domains are also needed. Accuracy requirements for small domains may be taken into account at the sampling design construction stage using suitable stratification, avoiding big clusters, and using compromise sample allocation between proportional allocation and equal allocation [20]. Unfortunately, it is not always possible to construct a stratified sample design with estimation domains as separate strata having a predetermined sample size, ensuring the accuracy of parameter estimates needed for these strata, because the whole sample size may become unacceptably large and a sampling design may become inefficient. Therefore, a sampling design is being constructed in order to ensure the accuracy of estimates for the whole population and large domains only. Thus, the design-based estimates of the parameters in domains with a low average sample size have high variance, despite the fact that they are unbiased or approximately unbiased. A domain with small average sample size is called a small area if the domain data-based estimator of the parameter has an unacceptably high variance.

A task for a statistician arises to use other estimators, different from the direct design-based estimators, and obtain more accurate estimates of the parameters in small areas. These estimators are constructed using models and auxiliary information about the population elements or data from the neighboring areas and are called small-area estimators (SAE).

One of the ways to improve accuracy for domain estimates is to use design-based generalized regression and calibration estimators with implicit use of models. This possibility is reviewed in [15]. For domains with extremely small sample size or even without sample at all, model-based estimators are constructed. Explicit models used for SAE may be divided into two main groups, depending on the type of the object for which the model is constructed: individual element and aggregate of elements (e.g., area). They are called element-level models and area-level models, respectively. Auxiliary information may be used at the level of the object for which the model is constructed or at the higher aggregation level for any of these models.

An introduction to model-based small area estimation methods is presented in [11], and their developments are given in [17].

The basic area level model was suggested by Fay and Herriot [8] in 1979. It is estimated by the best linear unbiased prediction, and the predictor consists of the weighted sum of the design-based estimator and synthetic estimator, where data from the outside domain is used. The subsequent developments of the Fay and Herriot model are reviewed in [3]. For area-level models, and especially for element level models, auxiliary information on various aggregation levels and multilevel models may be used, as described in [4]. The most popular methods used for the estimation of such models are the empirical best linear unbiased predictor (EBLUP) for a parameter of interest, the empirical Bayes method, and the hierarchical Bayes method for the approximation of the posterior distribution of the parameter, given data. Many books are devoted to the Bayesian inference in general, for example, [9], [19], and [10] for Bayesian analysis in finite populations.

The Labor Force Survey small-area parameter estimation problem attracted attention of many statisticians. When analyzing Labor Force Survey data, a categorical variable indicating individual's participation in the labor market (employed, unemployed, not in the labor force) may be resolved into three bivariate study variables. For example, a variable defining the unemployment status obtains the value 1 for an individual unemployed according to the survey definition and the value 0 otherwise. The finite-population parameter of interest is usually total, meaning the number of the unemployed in the population or its domain; mean, meaning the unemployment fraction in the population or in the domain; and ratio of two totals: the size of the unemployed population and the size of the labor force, expressing the unemployment rate. The main types of area-level and element-level models for a bivariate study variable are presented in the book [17].

A number of statisticians were using area-level models for Labor Force Survey data in their studies. Torelli and Trevisani [23] have applied the hierarchical Bayesian estimation method to the area-level model for the area parameters of the Italian Labor Force data. They were estimating domain parameters for the bivariate study variable and adding a synthetic estimator as a covariate to the linking linear regression model. The same authors in their paper [22] have made an overview of the methods and models used for labor force survey estimates for small geographical domains of Italy. They also support the use of a spatial component in the model estimated by the hierarchical Bayes estimation method.

Application of area-level models and the hierarchical Bayes estimation method for the estimation of the unemployment size for small areas in Poland is presented in [14].

A comprehensive study of municipal unemployment fractions is made by Boonstra et al. [2]. They have used estimators based on several different versions of the area-level model and also estimators based on an element-level logistic regression model for binary data and estimated sampling variances. The estimators are applied to the data set based on the Dutch Labor Force Survey. The authors in their paper [1] discuss issues concerning the model choice, including the use of linear (mixed) models for binary variables and the use of posterior means instead of maximum likelihood estimates.

Another direction of studies is connected with the use of element-level models. Datta et al. [5] used element-level models to estimate unemployment rates for United States small areas. Farrell [7] has estimated local labor force participation rates using the element level model with both area-level and element-level covariates estimated using the hierarchical Bayes methodology and applied it to the data set, based on the 1950 United States Census data.

A rotating sample design is usually used for the Labor Force Survey, and each individual is participating repeatedly in repetitive surveys. Therefore, the models for the estimation of unemployment rates are being generalized including the time series approach and spatial components. This generalization is used in the papers of Datta et al. [5], You, Rao, and Gambino [25], Fabrizi [6], Klimanek [12], and others.

The aim of the paper

The area-level models with the area-level covariates are being studied with the aim to find a suitable estimation method for a fraction of the unemployed for the Lithuanian Labor Force Survey by simulation. Empirical best linear unbiased prediction and hierarchical Bayes analysis is used for estimation. The reason for this is the attractiveness of the Bayesian method, popularity of this method among statisticians when solving the small area estimation problem, and an impression that any model can be estimated by this method. We refer to [2, 24] for the estimation of sampling variance and to [23] for the inclusion of the synthetic estimator in the linear regression model. The authors compose some of the mentioned model elements and propose their solution to the problem, including imputation in the case of missing data in a small area. The R package LaplacesDemon [21] is used to carry out the simulation study.

2 Notation

Let us denote a finite population by $\mathcal{U} = \{1, 2, \dots, N\}$, indicating population elements by their labels. Let y be a study variable with values $y_1, y_2, \dots, y_N$ in the population. Suppose that our study variable is bivariate and

$$y_k = \begin{cases} 1 & \text{if } k \text{ is unemployed,} \\ 0 & \text{otherwise,} \end{cases}$$

$k = 1, 2, \dots, N$. Let $\mathcal{U}$ be divided into M nonintersecting big domains $\mathcal{U}_j$: $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2 \cup \cdots \cup \mathcal{U}_M$, $\mathcal{U}_l \cap \mathcal{U}_j = \emptyset$, $l \neq j$, of size M_j, $j = 1, 2, \dots, M$, $M_1 + \cdots + M_M = N$.

Let $\mathcal{D}_i$ denote the nonintersecting small areas of size N_i, $i = 1, 2, \dots, D$, such that their unions coincide with the large domains:

$$\begin{aligned} \mathcal{U}_1 &= \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{D_1}, \\ \mathcal{U}_2 &= \mathcal{D}_{D_1+1} \cup \cdots \cup \mathcal{D}_{D_1+D_2}, \\ &\cdots \\ \mathcal{U}_M &= \mathcal{D}_{D_1+\cdots+D_{M-1}+1} \cup \cdots \cup \mathcal{D}_{D_1+D_2+\cdots+D_M}. \end{aligned}$$

We see that $\mathcal{U}_1$ consists of D_1 small areas, $\mathcal{U}_2$ consists of D_2 small areas, $\dots$, $\mathcal{U}_M$ consists of D_M small areas, $\mathcal{U} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_D$, $D_1 + \cdots + D_M = D$, $N_1 + \cdots + N_D = N$.

Let $q_j = (1/M_j) \sum_{k \in \mathcal{U}_j} y_k$ be the fraction of the unemployed for a big domain $\mathcal{U}_j$, $j = 1, 2, \dots, M$, and $\theta_i = (1/N_i) \sum_{k \in \mathcal{D}_i} y_k$ be the fraction of the unemployed for a small area $\mathcal{D}_i$, $i = 1, 2, \dots, D$. It will be the parameter of interest in this study. Attention has to be paid to the fact that the fraction of the unemployed is a different parameter from the unemployment rate according to the International Labor Organization definition, usually estimated in national statistical agencies.

Let us suppose that $\mathbf{s} \subset \mathcal{U}$ is an n size sample, drawn from the finite population according to the simple random sampling design, with the random sample sizes n_i in the small domains, $n_1 + \cdots + n_D = n$, and sample sizes m_i in the big domains, $m_1 + \cdots + m_M = n$.

Let us denote by $\widehat{q}_i$, $\widehat{\theta}_i$ the Horvitz–Thompson estimators for the fractions of the unemployed in big domains and small domains, respectively:

$$\widehat{q}_j = \frac{1}{m_j} \sum_{k \in \mathcal{U}_j \cap \mathbf{s}} y_k, \qquad \widehat{\theta}_i = \frac{1}{n_i} \sum_{k \in \mathcal{D}_i \cap \mathbf{s}} y_k;$$

their variances, called sampling variances, are

$$\operatorname{Var}(\widehat{q}_j) = \left(\operatorname{E}\left(\frac{1}{m_j} \;\middle|\; m_j > 0\right) - \frac{1}{M_j}\right) s_j^{(1)2}, \quad s_j^{(1)2} = \frac{1}{M_j - 1} \sum_{k \in \mathcal{U}_j} (y_k - q_j)^2,$$

$$\operatorname{Var}(\widehat{\theta}_i) = \left(\operatorname{E}\left(\frac{1}{n_i} \;\middle|\; n_i > 0\right) - \frac{1}{N_i}\right) s_i^{(2)2}, \quad s_i^{(2)2} = \frac{1}{N_i - 1} \sum_{k \in \mathcal{D}_i} (y_k - \theta_i)^2,$$

and

$$\widehat{\operatorname{Var}}(\widehat{\theta}_i) = \left(1 - \frac{n_i}{N_i}\right) \frac{\widehat{s}_i^{(2)2}}{n_i}, \quad \widehat{s}_i^{(2)2} = \frac{1}{n_i - 1} \sum_{k \in \mathcal{D}_i \cap \mathbf{s}} (y_k - \widehat{\theta}_i)^2, \tag{2.1}$$

$j = 1, \dots M$, $i = 1, \dots, D$, is used as an estimator of $\operatorname{Var}(\widehat{\theta}_i)$. Let us suppose that the variances $\operatorname{Var}(\widehat{q}_j)$, $j = 1, 2, \dots, M$, are low enough, and the variances $\operatorname{Var}(\widehat{\theta}_i)$, $i = 1, 2, \dots, D$, are unacceptably high.

The parameters θ_i, $i = 1, 2, \dots, D$, are the objects of estimation of the current study, using the small-area estimation methods.

3 Models

We introduce the area-level models to estimate unemployment fractions in small areas.

3.1 Fay–Herriot model

Let $\mathbf{X} = (X_1, X_2, \dots, X_D)'$ denote a matrix of area-level L auxiliary variables with vectorial values $X_i = (x_{i1}, x_{i2}, \dots, x_{iL})'$ as information for the area $\mathcal{U}_i$ for $i = 1, 2, \dots, D$.

Let $\widehat{\theta}_1, \dots, \widehat{\theta}_D$ be the design-based Horvitz–Thompson estimators for small areas. The basic small-area model, introduced by Fay and Herriot [8] in 1979, consists of two equations, sampling model (3.1) and linking model (3.2):

$$\widehat{\theta}_i = \theta_i + e_i, \tag{3.1}$$

$$\theta_i = X_i' \boldsymbol{\beta} + v_i, \quad i = 1, 2, \dots, D. \tag{3.2}$$

We assume sampling errors e_i to be independently distributed with mean $\operatorname{E}_p(e_i \,|\, \theta_i) = 0$ and variances $\operatorname{Var}_p(e_i \,|\, \theta_i) = \psi_i$ (expectations are taken with respect to the sampling design p); due to the central limit theorem, sampling model (3.1) errors e_i, $i = 1, 2, \dots, D$, may be considered normally distributed. The parameters θ_i are related to area-specific auxiliary data $\mathbf{X}$ through a linear linking model. Here $\boldsymbol{\beta} = (\beta_1, \dots, \beta_L)'$ is an L-dimensional vector of linear regression model parameters, area-specific random effects v_i are independent, identically distributed, with means $\operatorname{E}_m v_i = 0$ and variances $\operatorname{Var}_m v_i = \sigma_v^2$ with respect to the model distribution m; we suppose that v_i are normally distributed.

The sampling variances ψ_i, $i = 1, \dots, D$, are unknown, they have been estimated, as suggested in [2], by

$$\widehat{\psi}_i = \left(1 - \frac{n_i}{N_i}\right) \frac{\widehat{\sigma}_e^2}{n_i},$$

with the pooled population variance σ_e^2 estimated for unbalanced data by

$$\widehat{\sigma}_e^2 = \frac{1}{n - D} \sum_{i=1}^{D} \sum_{k \in \mathcal{D}_i \cap \mathbf{s}} (y_k - \widehat{\theta}_i)^2,$$

where n is the number of elements in the sample.

Considering $\widehat{\psi}_i$ as known sampling variances, the empirical best linear unbiased predictor of θ_i is obtained in [8]:

$$\widehat{\theta}_i^{FH}(\widehat{\sigma}_v^2) = \widehat{\gamma}_i\widehat{\theta}_i + (1-\widehat{\gamma}_i)X_i'\widehat{\boldsymbol{\beta}}(\widehat{\sigma}_v^2), \quad i = 1, 2, \dots, D,$$

with

$$\widehat{\gamma}_i = \frac{\widehat{\sigma}_v^2}{\widehat{\psi}_i + \widehat{\sigma}_v^2}, \qquad \widehat{\boldsymbol{\beta}}(\widehat{\sigma}_v^2) = \left(\sum_{k=1}^{D}\widehat{\gamma}_k X_k X_k'\right)^{-1}\left(\sum_{k=1}^{D}\widehat{\gamma}_k X_k\widehat{\theta}_k\right).$$

Its mean square error is estimated by

$$\begin{aligned}
\widehat{\mathrm{MSE}}\big(\widehat{\theta}_i^{FH}(\widehat{\sigma}_v^2)\big) &= g_{1i}\big(\widehat{\sigma}_v^2\big) - \widehat{B}\big(\widehat{\sigma}_v^2\big)(1-\widehat{\gamma}_i)^2 + g_{2i}\big(\widehat{\sigma}_v^2\big) + g_{3i}\big(\widehat{\sigma}_v^2\big), \qquad (3.3)\\
g_{1i}\big(\widehat{\sigma}_v^2\big) &= \widehat{\gamma}_i\widehat{\psi}_i,\\
\widehat{B}\big(\widehat{\sigma}_v^2\big) &= 2\widehat{\sigma}_v^2\frac{D\sum_{k=1}^{D}\widehat{\gamma}_k^2 - \big(\sum_{k=1}^{D}\widehat{\gamma}_k\big)^2}{\big(\sum_{k=1}^{D}\widehat{\gamma}_k\big)^3},\\
g_{2i}\big(\widehat{\sigma}_v^2\big) &= (1-\widehat{\gamma}_i)^2 X_i'\left(\sum_{k=1}^{D}\frac{1}{\widehat{\sigma}_v^2+\widehat{\psi}_k}X_k X_k'\right)^{-1}X_i,\\
g_{3i}\big(\widehat{\sigma}_v^2\big) &= \frac{2D\widehat{\psi}_i^2}{(\widehat{\sigma}_v^2+\widehat{\psi}_i)^3}\left(\sum_{k=1}^{D}\frac{1}{\widehat{\sigma}_v^2+\widehat{\psi}_k}\right)^{-2}.
\end{aligned}$$

This is a widely used accuracy estimator, applied before also in [13]. It underestimates the mean square error because the estimation of sampling variances ψ_i is ignored and the variability of $\widehat{\psi}_i$ is not taken into account.

3.2 Logit-normal-normal linear area-level model

The basic area-level model (3.1), (3.2) does not allow a straightforward application of the hierarchical Bayes estimation method in our case because the fraction of the unemployed θ_i is a bounded value, $\theta_i \in (0,1)$, but the distribution of the sampling error e_i is unbounded. The logit transformation for the parameter θ_i is used:

$$z_i = \mathrm{logit}(\theta_i) = \ln\frac{\theta_i}{1-\theta_i}, \quad i = 1, 2, \dots, D.$$

The sampling model (3.1) is transformed to the following one:

$$\widehat{z}_i = \mathrm{logit}(\widehat{\theta}_i) = z_i + e_i', \qquad (3.4)$$

assuming that the sampling errors e_i' are independent with $\mathrm{E}_p(e_i' \,|\, z_i) = 0$, $\mathrm{Var}_p(e_i' \,|\, z_i) = \sigma_i^2,\ i = 1, \dots, D$, with the modified version of the basic area-level model and the linking model

$$z_i = X_i'\boldsymbol{\beta} + \delta_i, \qquad (3.5)$$

assuming that area specific random effects δ_i are independent and $\mathrm{E}_m\delta_i = 0$, $\mathrm{Var}_m\,\delta_i = \sigma_v^2$ with respect to the linking model.

Here we have an unmatched sampling and linking area-level model with unknown variances σ_i^2 of errors e_i' in the new sampling model (3.4). The hierarchical Bayes approach takes account of the uncertainties associated with unknown parameters in the model, as it has been done in [24].

For Bayesian analysis, we have to supplement the model (3.4), (3.5) with the assumptions about prior distributions of the parameters. Let $\boldsymbol{\beta}$ and σ_v^2 be independent, and $\boldsymbol{\beta}$, σ_v^2, σ_i^2, $i = 1, 2, \ldots, D$, have prior distributions close to "flat" distributions:

$$\begin{aligned} \beta_l &\sim \mathcal{N}(0, 1000), \quad && l = 1, 2, \ldots, L, \\ \sigma_v^{-2} &\sim \Gamma(a_0, b_0), && a_0 > 0,\ b_0 > 0, \\ \sigma_i^{-2} &\sim \Gamma(a_i, b_i), && a_i > 0,\ b_i > 0,\ i = 1, 2, \ldots, D. \end{aligned} \tag{3.6}$$

The values of the inverse gamma distribution parameters used here are $a_0 = b_0 = 0.01$, $a_i = b_i = 0.01$, $i = 1, 2, \ldots, D$.

For any domain, it may happen that there is no sample in it. Let us use Bayes imputation:

$$\widehat{z}_i \to \widehat{z}_{i\,\mathrm{imputed}} = \begin{cases} \widehat{z}_i & \text{for } \widehat{z}_i \text{ present}, \\ \alpha \sim \mathcal{N}(\alpha_0, \alpha_1) & \text{for } \widehat{z}_i \text{ missing}, \end{cases} \tag{3.7}$$

with random parameters α_0, α_1 independent of $\boldsymbol{\beta}$, σ_v^2, $\sigma_1^2, \ldots, \sigma_D^2$. The parameters α_0, α_1 of the random variable α will be estimated in Bayesian analysis with their prior distributions $\mathcal{N}(0, 1000)$.

3.2.1 Application of hierarchical Bayes analysis

The aim of Bayesian analysis is to obtain the posterior distribution $f(\theta_i \,|\, \widehat{\theta}_i)$ of small-area parameters, given the Horvitz–Thompson estimator $\widehat{\theta}_i$ of this area, called here "data," and the subjective prior distribution $f(\lambda)$ of model parameters for $\lambda = (\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_D^2, \sigma_v^2, \alpha_1, \alpha_2)$.

Applying the Bayes theorem, we obtain the joint density of estimation parameters and model parameters [19]:

$$f(\theta, \lambda \,|\, \widehat{\theta}) = \frac{f(\widehat{\theta}, \theta \,|\, \lambda) f(\lambda)}{f_1(\widehat{\theta})}$$

for $\theta = (\theta_1, \ldots, \theta_D)'$ and $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_D)'$ with a marginal density of $\widehat{\theta}$ in the denominator:

$$f_1(\widehat{\theta}) = \int f(\widehat{\theta}, \theta \,|\, \lambda) f(\lambda)\, \mathrm{d}\theta\, \mathrm{d}\lambda.$$

The desired posterior density is obtained as

$$f(\theta \,|\, \widehat{\theta}) = \int f(\theta, \lambda \,|\, \widehat{\theta})\, \mathrm{d}\lambda = \int f(\theta \,|\, \widehat{\theta}, \lambda) f(\lambda \,|\, \widehat{\theta})\, \mathrm{d}\lambda, \tag{3.8}$$

from which we find $\mathrm{E}(\theta \,|\, \widehat{\theta})$ and $\mathrm{Var}(\theta \,|\, \widehat{\theta})$, and they are considered as estimates of the parameters of interest and the estimates of their variances, respectively.

Because of difficulties in the integration of (3.8), this integral is approximated using the Markov chain Monte Carlo method. The essence of the method is the construction of a Markov chain such that its univariate distribution converges to a unique stationary distribution, the density of which equals $f(\theta \,|\, \widehat{\theta})$.

For easier realization of the Markov chain, a Gibbs sampler is used, which means a process of iterative conditional sampling of parameters from the joint distribution of parameters λ. Difficulties in this sampling are circumvented by a griddy Gibbs sampler. Its essence is that, instead of sampling from the conditional distribution performing the Gibbs sampler, the conditional distribution density functions of parameters $f(\lambda \,|\, \widehat{\theta})$ are evaluated on a sequence of grid points, and an approximation to the conditional distribution by a discrete empirical distribution function is obtained.

The griddy Gibbs sampler for hierarchical Bayes analysis is introduced in [18]. This procedure is also described in detail in [7] when analyzing the element-level model for proportion; it has been also used in [16].

The hierarchical Bayes method is used to estimate unmatched model (3.4)–(3.7) for $\widehat{z}_i = \operatorname{logit}(\widehat{\theta}_i)$. Let $\theta_{i1}, \dots, \theta_{iK}$, $i = 1, 2, \dots, D$, denote K independent draws from the posterior distribution $f(\theta \,|\, \widehat{\theta})$, obtained by hierarchical Bayesian analysis. The posterior mean $\mathrm{E}(\theta \,|\, \widehat{\theta})$ and posterior variance $\mathrm{Var}(\theta \,|\, \widehat{\theta})$ are estimated by

$$\bar{\theta}_i = \frac{1}{K} \sum_{k=1}^{K} \theta_{ik} \quad \text{and} \quad \widehat{\mathrm{Var}}(\theta_i) = \frac{1}{K-1} \sum_{k=1}^{K} (\theta_{ik} - \bar{\theta}_i)^2, \tag{3.9}$$

respectively, $i = 1, 2, \dots, D$, and they are used as estimates of small-area parameters and accuracy measures of estimates.

3.2.2 Logit-normal-normal area-level model with the synthetic estimator as covariate

If a small area $\mathcal{D}_i$ is a subset of a large area $\mathcal{U}_j$, then it may be expected that a fraction θ_i should be around the value q_j. Considering the estimate $\widehat{q}_j$ as accurate enough, let us add the term $\operatorname{logit}(\widehat{q}_j)$ to the mean of the distribution of $z_i = \operatorname{logit}(\theta_i)$ in a linking model (3.5) and obtain the new model:

$$\begin{aligned}
\widehat{z}_i \,|\, \theta_i &\sim \mathcal{N}\big(z_i, \sigma_i^2\big), \qquad z_i \,|\, \boldsymbol{\beta},\, \sigma_v^2 \sim \mathcal{N}\big(\operatorname{logit}(\widehat{q}_j) + X_i'\boldsymbol{\beta}, \sigma_v^2\big), \\
\beta_l &\sim \mathcal{N}(0, 1000), \quad l = 1, 2, \dots, L, \\
\sigma_v^{-2} &\sim \Gamma(a_0, b_0), \quad \sigma_i^{-2} \sim \Gamma(a_i, b_i), \quad a_0, b_0, a_i, b_i > 0, \\
\widehat{z}_i \to \widehat{z}_{i\,\mathrm{imputed}} &= \begin{cases} \widehat{z}_i & \text{for } \widehat{z}_i \text{ present,} \\ \alpha \sim \mathcal{N}(\alpha_0, \alpha_1) & \text{for } \widehat{z}_i \text{ missing,} \end{cases}
\end{aligned}$$

for the ith small area (municipality) belonging to the jth larger domain (county), $i = 1, 2, \dots, D$, $j = 1, 2, \dots, M$.

The parameters $\sigma_1^2, \dots, \sigma_D^2$, $\boldsymbol{\beta}$, σ_v^2, α_0, α_1 are supposed to be independent, with prior distribution $\mathcal{N}(0, 1000)$ for α_0, α_1. The Markov chain Monte Carlo algorithm and the griddy Gibbs sampler, as before, are used for the approximation of the posterior distribution $f(\theta \,|\, \widehat{\theta})$.

3.3 Binomial-logit-normal model

The number of the unemployed belonging to the sampled part of a small domain $\mathcal{D}_i$ may be expressed as a sum of values of variable y:

$$t_i = \sum_{k \in \mathcal{D}_i \cap \mathbf{s}} y_k, \quad i = 1, 2, \dots, D.$$

The random variable t_i is distributed according to the binomial distribution with unknown parameter θ_i. For this parameter, we assume a linear regression model to be a linking model with area-specific random effects δ_i independent and identically normally distributed. Assumptions concerning prior distributions are made, and the model is as follows:

$$\begin{aligned}
t_i \,|\, \theta_i &\sim \operatorname{binomial}(n_i, \theta_i), \\
z_i &= \operatorname{logit}(\theta_i) = X_i'\boldsymbol{\beta} + \delta_i, \\
\delta_i &\sim \mathcal{N}(0, \sigma_v^2), \qquad \beta_l \sim \mathcal{N}(0, 1000), \\
\sigma_v^{-2} &\sim \Gamma(a, b), \quad a > 0,\ b > 0,
\end{aligned} \tag{3.10}$$

where $\boldsymbol{\beta}$ and σ_v^2 are assumed to be mutually independent, $i = 1, 2, \dots, D$, $l = 1, 2, \dots, L$.

Imputation is applied, if needed:

$$\widehat{z}_i \to \widehat{z}_{i\,\mathrm{imputed}} = \begin{cases} \widehat{z}_i & \text{for } \widehat{z}_i \text{ present,} \\ \mathrm{logit}(\widehat{q}_j) & \text{for } \widehat{z}_i \text{ missing,} \end{cases}$$

for ith small area belonging to the jth larger domain. The model is estimated by the same method: Markov chain Monte Carlo and griddy Gibbs sampler.

Binomial-logit-normal model with synthetic estimates included among covariates

Here the model of Section 3.3 is used with the linking model (3.10) replaced by

$$z_i = \mathrm{logit}(\theta_i) = \mathrm{logit}(\widehat{q}_j) + X_i'\boldsymbol{\beta} + \delta_i$$

for ith small area belonging to the larger domain $\mathcal{U}_j$, in the same way as it has been done in Section 3.2.2.

4 Accuracy measures

In order to compare the accuracy of the models suggested, R samples are drawn from the finite population, and estimates of the fraction of the unemployed are calculated according to all the models described.

Let $\bar{\theta}_{ir}$, $i = 1, 2, \ldots, D$, denote the small-area estimates $\bar{\theta}_i$ (3.9) in the rth simulation run, $r = 1, 2, \ldots, R$. The following accuracy measures are used:

1. Root mean square error

$$\mathrm{rmse}_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\bar{\theta}_{ir} - \theta_i)^2}, \quad \mathrm{rmse} = \frac{1}{D}\sum_{i=1}^{m}\mathrm{rmse}_i.$$

2. Relative root mean square error

$$\mathrm{rmseREL}_i = \frac{\mathrm{rmse}_i}{\theta_i}, \quad \mathrm{rmseREL} = \frac{1}{D}\sum_{i=1}^{D}\mathrm{rmseREL}_i.$$

3. Simulation standard error

$$\mathrm{seSIM}_i = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\bar{\theta}_{ir} - \frac{1}{R}\sum_{k=1}^{R}\bar{\theta}_{ik}\right)^2}, \quad \mathrm{seSIM} = \frac{1}{D}\sum_{i=1}^{D}\mathrm{seSIM}_i.$$

4. Simulation bias

$$\mathrm{biasSIM}_i = \frac{1}{R}\sum_{r=1}^{R}\bar{\theta}_{ir} - \theta_i, \quad \mathrm{biasSIM} = \frac{1}{D}\sum_{i=1}^{D}\mathrm{biasSIM}_i.$$

5. Root mean simulation mean squared error estimates or posterior variances

$$\mathrm{seEST}_i = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\widehat{\mathrm{Var}}_r(\theta_i)}, \quad \mathrm{seEST} = \frac{1}{D}\sum_{i=1}^{D}\mathrm{seEST}_i.$$

   Here $\widehat{\mathrm{Var}}_r(\theta_i)$ denotes the estimate of variance (2.1) in the case of Horvitz–Thompson estimator, the estimate of mean squared error (3.3) in the case of Fay–Herriot model based estimate, and the estimate of posterior variance (3.9) in the case of Bayes analysis in the rth simulation run.

The estimates presented are compared with respect to these measures.

5 Simulation study

Lithuanian Labor Force Survey data of the 4th quarter of 2012 is used for simulation.

The population consists of $N = 22\,382$ individuals, 15–74 years old. The sample size used is $n = 2\,000$ individuals. Auxiliary variables selected for a fraction of the unemployed by stepwise regression are small-area fractions of

- registered unemployed individuals,
- males,
- urban population,
- 55–74 year old population.

They are considered as known. The linear regression model for the Horvitz–Thompson estimates of the fraction of the unemployed on the population level, based on the auxiliary variables mentioned, explains 34% of the total sum of squares. The population is divided into $M = 10$ big domains, counties, and counties are divided into $D = 57$ small domains, municipalities.

Notation for estimators of small-area fractions θ_i, $i = 1, 2, \ldots, D$:

HT – Horvitz–Thompson estimator,
FH – EBLUP Fay–Herriot model-based estimator,
LNN1 – unmatched logit-normal-normal model-based estimator,
LNN2 – unmatched logit-normal-normal with synthetic covariate model-based estimator,
BLN1 – binomial-logit-normal model-based estimator,
BLN2 – binomial-logit-normal with synthetic covariate model-based estimator.

$R = 30$ simulation runs are implemented in a design-based simulation study. Some of the estimates obtained are presented in Fig. 1. We can observe a smoothing effect of the Fay–Herriot model-based estimates.

The small areas are classified into three classes, each consisting of 19 areas, by their size. The domain size intervals and corresponding sample size intervals, obtained by simulation, are presented in Table 1. The behavior of accuracy measures of estimates depending on the domain size shown in Table 2. The summary of accuracy measures for all domains is presented in Table 3.

A computer program LaplacesDemon [21] is used for estimation. Unfortunately, the use of the griddy Gibbs sampler increases the simulation time dramatically. Estimation of variances in sampling model by the hierarchical Bayes method makes estimation process extremely time consuming.

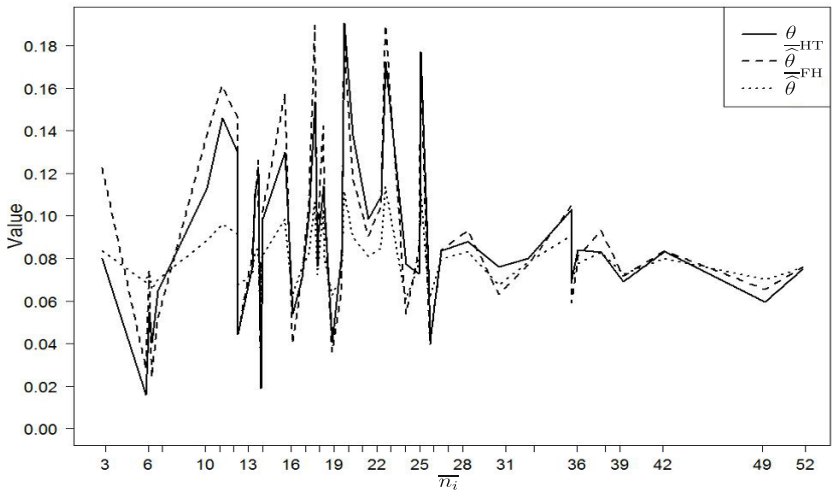

Figure 1. Estimates for the fraction of the unemployed.

Table 1. Construction of area size classes

Domain size class	Domain size	Sample size
Small	(25, 185)	(3, 17)
Average	(188, 289)	(18, 26)
Large	(293, 3617)	(27, 324)

Table 2. Accuracy measures of the estimates by area size classes

Domain size class	Estimator					
	HT	FH	LNN1	LNN2	BLN1	BLN2
	rmse					
Small	0.0108	0.0015	0.0029	0.0032	0.0022	0.0026
Average	0.0042	0.0019	0.0020	0.0024	0.0018	0.0021
Large	0.0019	0.0006	0.0009	0.0016	0.0004	0.0008
	rmseREL					
Small	1.3781	0.6912	1.1356	1.1008	0.9594	1.0298
Average	0.6417	0.3795	0.4791	0.5088	0.3987	0.4456
Large	0.5179	0.3120	0.3891	0.4949	0.2365	0.3469
	seSIM					
Small	0.0922	0.0277	0.0250	0.0347	0.0163	0.0293
Average	0.0624	0.0276	0.0220	0.0290	0.0119	0.0271
Large	0.0405	0.0230	0.0162	0.0258	0.0082	0.0219
	biasSIM					
Small	0.0085	−0.0031	0.0310	0.0328	0.0236	0.0218
Average	0.0003	−0.0214	0.0128	0.0094	0.0000	0.0005
Large	−0.0019	−0.0029	0.0214	0.0225	0.0104	0.0082
	seEST					
Small	0.0071	0.0018	0.0059	0.0069	0.0041	0.0042
Average	0.0044	0.0012	0.0051	0.0059	0.0030	0.0032
Large	0.0015	0.0007	0.0032	0.0043	0.0011	0.0014

Table 3. Average accuracy measures of the estimates

Estimator	HT	FH	LNN1	LNN2	BLN1	BLN2
rmse	0.0056	**0.0014**	0.0019	0.0024	0.0015	0.0018
rmseREL	0.8459	**0.4609**	0.6680	0.7015	0.5316	0.6075
seSIM	0.0650	0.0261	0.0211	0.0299	**0.0121**	0.0261
biasSIM	**0.0023**	−0.0091	0.0217	0.0215	0.0113	0.0102
seEST	0.0043	**0.0012**	0.0047	0.0057	0.0027	0.0029

Conclusions. Simulation results show that:

- All accuracy measures of all estimators are increasing with decreasing domain size and sample size.
- All accuracy measures of the estimates, except of simulated bias obtained by modeling and small-area estimation methods, are lower than in the case of the Horvitz–Thompson estimator.
- Model-based estimates may have some simulation bias, which is sometimes higher than that for the Horvitz–Thompson estimator. The bias is usually higher for small areas than for large areas.
- The Fay–Herriot model estimated using empirical best linear unbiased prediction has the smallest relative root mean squared error and simulation standard error.

- The relative root mean square error is higher for the log-normal-normal model than for the Fay–Herriot model because of the estimation of sampling variances for the first one.
- The inclusion of a synthetic component in the models increases the root mean square error for the logit-normal-normal model and for the binomial-logit-normal model.

Comparison of accuracy of model-based estimators and model-assisted estimators [15] for the fraction of the unemployed in small areas should be done by simulation in future.

6 Discussion

Attention has to be paid to the fact that the mean square error of the Fay–Herriot model is underestimated because the estimated sampling variance is kept there as known and the variation of sampling variance estimates is not taken into account. One has to be cautious in giving priority to this estimator. You and Chapman [24] have shown that the underestimation of the mean squared error $\mathrm{MSE}(\widehat{\theta})$ for the Fay–Herriot model may be significant in the case of a small domain sample size and insignificant for a high domain sample size.

Usually, estimates in small areas have to add up to higher aggregation level estimates in official statistics. It will not be so for model-based estimates. Therefore, the sum of small-area estimates has to be benchmarked in some way. It may change the values of some accuracy measures a little, but this step is unavoidable if the harmonized system of estimates is needed.

Usually, for the surveys of national statistical agencies, the usage of a common weighting system for all estimates is needed. Actually, it is important only for big domain estimates, which are used for government needs and international comparison. Calibration estimators and regression estimators may be used to improve the accuracy of those design-based estimates. For small-area estimates that are obtained for local needs, carefully chosen model-based estimators and benchmarking may be used if assumptions for model application are satisfied.

Sample designers should establish the desired degree of precision not only for national level estimates but also for domains of interest. Design-based model-assisted and especially model-based estimators are efficient and may reduce the sampling error if good auxiliary variables are available. Bias of the estimates is possible if model assumptions are not satisfied. Model-based estimators should be used with caution even if they have significantly smaller coefficients of variation.

Acknowledgment. The authors are thankful to the anonymous referee for valuable comments, resulting in a significant improvement of the paper.

References

1. H.J. Boonstra, B. Buelens, K. Leufkens, and M. Smeets, Small area estimates of labour status in Dutch municipalities, Discussion paper No. 201102, Statistics Netherlands, The Hague, 2011.

2. H.J. Boonstra, B. Buelens, and M. Smeets, Estimation of municipal unemployment fractions – a simulation study comparing different small area estimators, Project No. DMH-205714, Statistics Netherlands, The Hague, 2009.

3. G. Datta and M. Ghosh, Small area shrinkage estimation, *Stat. Sci.*, **27**(1):95–114, 2012.

4. G.S. Datta, Model-based approach to small area estimation, in D. Pfeffermann and C.R. Rao (Eds.), *Handbook of Statistics. Sample Surveys: Inference and Analysis, Vol. 29B*, Elsevier, North Holland, 2009, pp. 251–288.

5. G.S. Datta, P. Lahiri, T. Maiti, and K.L. Lu, Hierarchical Bayes estimation of unemployment rates for the states of the U.S., *J. Am. Stat. Assoc.*, **94**(448):1074–1082, 1999.

6. E. Fabrizi, Hierarchical Bayesian models for the estimation of unemployment rates in small domains of the Italian Labour Force Survey, *Statistica*, **LXII**(4):603–618, 2002.

7. P.J. Farrell, Bayesian inference for small area proportions, *Sankhyā, Ser. B*, **62**(3):402–416, 2000.

8. R.E. Fay and R.A. Herriot, Estimates of income for small places: An application of James–Stein procedures to census data, *J. Am. Stat. Assoc.*, **74**(366):269–277, 1979.

9. J. Geweke, *Contemporary Bayesian Econometrics and Statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, 2005.

10. M. Ghosh and G. Meeden, *Bayesian Methods for Finite Population Sampling*, Chapman & Hall, London, 1997.

11. M. Ghosh and J.N.K. Rao, Small area estimation: An appraisal, *Stat. Sci.*, **9**(1):55–93, 1994.

12. T. Klimanek, Using indirect estimation with spatial autocorellation in social surveys in Poland, *Przegl. Stat.*, **1**(spec. issue):155–172, 2012.

13. D. Krapavickaitė, An example of small area estimation in finite population sampling, *Liet. Mat. Rink.*, **43**(spec. issue):497–503, 2003.

14. J. Kubacki, Application of the hierarchical Bayes estimation to the Polish Labour Force Survey, *Statistics in Transition*, **6**(5):785–796, 2004.

15. R. Lehtonen and A. Veijanen, Design-based methods of estimation for domains and small areas, in D. Pfeffermann and C.R. Rao (Eds.), *Handbook of Statistics. Sample Surveys: Inference and Analysis*, Volume 29B, Elsevier, North Holland, 2009, pp. 219–249.

16. D. Malec, J. Sedransk, C.L. Moriarity, and F.B. LeClere, Small area inference for binary variables in the national health interview survey, *J. Am. Stat. Assoc.*, **92**(439):815–826, 1997.

17. J.N.K. Rao, *Small Area Estimation*, John Wiley & Sons, Hoboken, NJ, 2003.

18. C. Ritter and M. A. Tanner, Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler, *J. Am. Stat. Assoc.*, **87**(419):861–868, 1992.

19. C.P. Robert, *The Bayesian Choice*, Springer, New York, 2007.

20. M.P. Singh, J. Gambino, and H. Mantel, Issues and strategies for small area data, *Survey Methodology*, **20**(1):3–22, 1994.

21. Statisticat, LLC, Laplacesdemon: Complete environment for Bayesian inference, R package version 14.04.05, 2014, available from: `http://www.bayesian-inference.com/software`.

22. N. Torelli and M. Trevisani, Labour force estimates for small geografical domains in Italy: Problems, data and models, Working paper No. 118, University of Trieste, 2008.

23. M. Trevisani and N. Torelli, Hierarchical Bayesian models for small area estimation with count data, Working paper No. 115, University of Trieste, 2007.

24. Y. You and B. Chapman, Small area estimation using area level models and estimated sampling variances, *Survey Methodology*, **32**(1):97–103, 2006.

25. Y. You, J.N.K. Rao, and J. Gambino, Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach, *Survey Methodology*, **29**(1):25–32, 2003.