A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 252, 2011

Tomasz Żądło[*]

ON SOME CALIBRATION ESTIMATORS OF SUBPOPULATION TOTAL FOR LONGITUDINAL DATA

Abstract. The problem of modeling longitudinal profiles is considered assuming that the population and elements affiliation to subpopulation may change in time. The considerations are based on a model with auxiliary variables for longitudinal data with subject specific (in this case - element and subpopulation specific) random components (compare Verbeke, Molenberghs, 2000; Hedeker, Gibbons, 2006) which is a special case of the General Linear Mixed Model. In the paper calibration estimators of subpopulation total for data from one period are presented and some modifications for the case of longitudinal data are proposed. Design-based mean squared errors and its estimators are also presented. In the simulation study accuracy of the estimators is compared with Horvitz-Thomson estimator and the best empirical linear unbiased predictor derived for the considered model.

Keywords: *longitudinal data; general linear mixed model; empirical best linear unbiased predictor; calibration estimators.*

1. SMALL AREA ESTIMATION AND LONGITUDINAL SURVEYS

In survey sampling, the problem of estimation or prediction of subpopulations' (domains') characteristics has become a very important issue. Besides, in the case of longitudinal surveys it is possible to increase the accuracy of the estimators or predictors by using information from other periods or even to estimate or predict subpopulation's characteristic for a period when the number of sampled domain elements equals zero. Domains with small or zero sample sizes are called small areas. The proposed solutions can be used by opinion polls companies, the market research sector and statistical offices during surveys conducted on behalf of different types of enterprises, local authorities or even the central government to obtain information that is useful or even essential in making decisions about, *inter alia*, fund allocation, investments, health care or environmental protection.

[*] PhD., Department of Statistics, University of Economics In Katowice.

2. CALIBRATION ESTIMATOR OF POPULATION TOTAL FOR DATA FROM ONE PERIOD

Based on the design approach, Deville and Särndal [1992] propose an estimator of population total given by:

$$\hat{t}^{CAL} = \sum_{i \in s} w_{si} y_i \,, \tag{1}$$

where: s is a sample of size n drawn from a population Ω of size N, weights w_{si} fulfill a so-called calibration equation. Deville and Särndal [1992] propose a calibration equation given by:

$$\forall_{k \in \{1,2,\ldots,p\}} \sum_{i \in s} w_{si} x_{ik} = \sum_{i \in \Omega} x_{ik} \,, \tag{2}$$

where: p is the number of auxiliary variables. Based on a matrix formula, equation (2) may be written as:

$$\sum_{i \in s} w_{si} \mathbf{x}_i = \sum_{i \in \Omega} \mathbf{x}_i \,,$$

where $\mathbf{x}_i^T = \left[x_{i1} \quad x_{i2} \quad \ldots \quad x_{ip} \right]$.

In the case of the design approach weights giving perfect estimates for auxiliary variables should – intuitively – increase estimation accuracy for the variable of interest. Deville and Särndal [1992] argue that "... weights that perform well for the auxiliary variable should perform well for the study variable". On the other hand, the calibration equation is a condition of model-unbiasedness of the population total predictor under the General Linear Model (GLM). Moreover, Deville and Särndal [1992] study the problem of design-unbiasedness of the estimator looking for weights fulfilling the calibration equation and conditionally minimizing some distance measure from the basic design weights $d_i = \pi_i^{-1}$ (where π_i are first order inclusion probabilities). For example, the following quadratic distance measure is considered:

$$\psi^{(s)} = \sum_{i \in s} \frac{(w_{si} - d_i)^2}{d_i q_i} \,, \tag{3}$$

where:

w_{si} - weights of (1) fulfilling condition (2),

q_i - some positive weights uncorrelated with d_i (introduced to obtain a more general solution).

A design-consistent calibration estimator obtained by conditional minimization of (3) is called the GREG estimator (the generalized regression estimator)

and it will be denoted by $\hat{t}^{GREG}$. Weights obtained by conditional minimization of (3) are the following[1]:

$$w_{si} = g_{si} d_i, \tag{4}$$

where:

$$g_{si} = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_x^{HT})^T \left(\sum_{i \in s} d_i q_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i q_i, \tag{5}$$

$\mathbf{t}_x = \sum_{i \in \Omega} \mathbf{x}_i$, $\hat{\mathbf{t}}_x^{HT} = \sum_{i \in s} d_i \mathbf{x}_i$.

Using weights (4), we obtain an estimator given by:

$$\hat{t}^{GREG} = \sum w_{si} y_i = \hat{t}_y^{HT} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x^{HT} \right)^T \hat{\mathbf{B}}, \tag{6}$$

where:

$$\hat{t}_y^{HT} = \sum_{i \in s} d_i y_i,$$

$$\hat{\mathbf{B}} = \left(\sum_{i \in s} d_i q_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i \in s} d_i q_i \mathbf{x}_i y_i. \tag{7}$$

Deville and Särndal[2] prove that calibration estimators obtained by conditional minimization of different distance measures are asymptotically equivalent to the GREG estimator in the following sense:

$$N^{-1} \left(\hat{t}^{CAL} - \hat{t}^{GREG} \right) = O_p(n^{-1}). \tag{8}$$

Besides, simulation analyses[3] show that the values and accuracy of different calibration estimators (calibration estimators obtained by conditional minimization of different distance measures) are similar even for small sample sizes. Because the calibration equation may not be fulfilled Theberge [2000] proposes to consider the interval to which weights should belong to instead of the calibration equation.

To derive asymptotic design-variance (denoted by $\breve{D}_p^2(.)$) of the GREG estimator, Deville and Särndal [1992] use Taylor approximation and obtain the following formula:

---

[1] For example: Särndal C.E., Swensson B., Wretman J., [1992], *Model assisted survey sampling*, Springer-Verlag, New York, p.232; Rao J.N.K., [2003], *Small area estimation*. John Wiley & Sons, New York, p.13.

[2] Deville J.C., Särndal C.E., [1992], *Calibration estimators in survey sampling*, Journal of the American Statistical Association, 87, p.379.

[3] For example: Singh A.C., Mohl C.A., [1996], *Understanding calibration estimators in survey sampling*, Survey methodology, 22, pp.107-115; Stukel D.M., Hidiroglou M.A., Särndal C.E., [1996], *Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization,* Survey methodology, 22, 177-125.

$$\breve{D}_p^2\left(\hat{t}^{GREG}\right)=\sum_{i\in\Omega}\sum_{\substack{j\in\Omega\\ j>i}}\left(\pi_{ij}-\pi_i\pi_j\right)\left(\frac{E_i}{\pi_i}-\frac{E_j}{\pi_j}\right)^2, \tag{9}$$

where π_{ij} are second order inclusion probabilities, $E_i=y_i-\mathbf{x}_i^T\mathbf{B}$, and $\mathbf{B}$ is a solution of $\left(\sum_{i\in\Omega}q_i\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{B}=\sum_{i\in\Omega}q_i\mathbf{x}_i\mathbf{x}_i^T$.

Furthermore, Deville and Särndal[4] suggest to use (9) for other calibration estimators because of (8). To estimate (9), the following *p*-consistent (e.g. Rao[5]) Sen-Yates-Grundy's type of estimator may be used:

$$\hat{D}_p^2\left(\hat{t}^{GREG}\right)=\sum_{i=1}^{n}\sum_{\substack{j=1\\ j>i}}^{n}\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{e_i}{\pi_i}-\frac{e_j}{\pi_j}\right)^2, \tag{10}$$

where:

$$e_i=y_i-\mathbf{x}_i^T\hat{\mathbf{B}}. \tag{11}$$

and $\hat{\mathbf{B}}$ is given by (7). Because estimator (10) may underestimate the variance of the GREG estimator, another *p*-consistent estimator may be used (Rao[6])

$$\hat{D}_p^2\left(\hat{t}^{GREG}\right)=\sum_{i=1}^{n}\sum_{\substack{j=1\\ j>i}}^{n}\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{g_{si}e_i}{\pi_i}-\frac{g_{si}e_j}{\pi_j}\right)^2, \tag{12}$$

where g_{si} is given by (5). This estimator will be taken into consideration in the simulation analysis.

3. MODEL-CALIBRATION ESTIMATOR OF POPULATION TOTAL FOR DATA FROM ONE PERIOD

Wu and Sitter [2001] and then Wu [2003] propose and study a model calibration estimator of population total for data from one period. Let $Y_1, Y_2, ..., Y_N$ be independent random variables (which will not be true for the superpopulation model considered in this paper) of some joint distribution ξ and let us assume that[7]:

$$\begin{cases} E_\xi(Y_i)=\mu(\mathbf{x}_i,\boldsymbol{\theta}) \\ D_\xi^2(Y_i)=v_i^2\sigma^2 \end{cases}, \tag{13}$$

[4] Deville J.C., Särndal C.E., [1992], p.379.
[5] Rao J.N.K., [2003], p.15.
[6] Ibidem.
[7] Wu C., Sitter R.R., [2001], *A model-calibration approach to using complete auxiliary information from survey data*, Journal of the American Statistical Association, 96, p.186.

where $i = 1,...,N$, $\boldsymbol{\theta} = \left[\theta_1\ \theta_2\ ...\ \theta_p\right]^T$ and σ^2 are unknown superpopulation parameters, $\mu(\mathbf{x}_i, \boldsymbol{\theta})$ is some (e.g. nonlinear) known function of $\mathbf{x}_i$ and $\boldsymbol{\theta}$, v_i is some known function of $\mathbf{x}_i$. Wu and Sitter (2001) consider a design-based estimator of model parameters $\boldsymbol{\theta}$ given by:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}_s^T \boldsymbol{\Pi}^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \boldsymbol{\Pi}^{-1} \mathbf{y}_s, \tag{14}$$

where $\mathbf{y}_s$ is a $n \times 1$ vector of values of the variable of interest, $\mathbf{X}_s$ is $n \times p$ matrix of auxiliary variables and $\boldsymbol{\Pi} = \mathrm{diag}_{1 \le i \le n}\left(\pi_i\right)$ is a diagonal matrix of first order inclusion probabilities.

Wu and Sitter [2001] propose a model-calibration estimator obtained by conditional minimization of the distance measure (3) subject to the following constraints

$$N^{-1} \sum_{i \in s} w_i = 1 \wedge \sum_{i \in s} w_{si} \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = \sum_{i \in \Omega} \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}). \tag{15}$$

They note that in the original formulation of the calibration estimator the constraint $N^{-1} \sum_{i \in s} w_i = 1$ is not present (although it can be introduced assuming that all values of one auxiliary variable equal one). It should be noted that the model-calibration equations (15) simplify to the classic calibration equation (2) in the case of a linear superpopulation model with population specific (but not domain specific) parameters.

The resulting model-calibration estimator is given by:

$$\hat{t}^{MCAL} = \hat{t}_y^{HT} + \left(\sum_{i \in \Omega} \hat{\mu}_i - \sum_{i \in s} d_i \hat{\mu}_i \right) \hat{\mathbf{B}}_N \tag{16}$$

where:

$$\hat{B}_N = \left(\sum_{i \in s} d_i q_i (\hat{\mu}_i - \overline{\mu})^2 \right)^{-1} \sum_{i \in s} d_i q_i (\hat{\mu}_i - \overline{\mu})(y_i - \overline{y}), \tag{17}$$

$\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$, $\overline{y} = \left(\sum_{i \in s} d_i q_i \right)^{-1} \sum_{i \in s} d_i q_i y_i$, $\overline{\mu} = \left(\sum_{i \in s} d_i q_i \right)^{-1} \sum_{i \in s} d_i q_i \hat{\mu}_i$.

Under some assumptions presented by Wu and Sitter[8] the asymptotic design-variance of $\hat{t}^{MC}$ is given by:

$$\hat{D}_p^2\left(\hat{t}^{MCAL}\right) = \sum_{i \in \Omega} \sum_{\substack{j \in \\ j > i}} \left(\pi_{ij} - \pi_i \pi_j\right) \left(\frac{U_i}{\pi_i} - \frac{U_j}{\pi_j} \right)^2, \tag{18}$$

[8] Ibidem, p.187.

where $U_i = y_i - \mu_i B_N$, $\mu_i = \mu(\mathbf{x}_i, \boldsymbol{\theta}_N)$, $\boldsymbol{\theta}_N = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, $\mathbf{X}$ is $N \times p$ matrix of auxiliary variables, $\mathbf{y}$ is a vector of N values of the variable of interest, $B_N = \left(\sum_{i\in\Omega} q_i (\hat{\mu}_i - \bar{\mu}_N)^2\right)^{-1} \sum_{i\in\Omega} q_i (\hat{\mu}_i - \bar{\mu}_N)(y_i - \bar{y}_N)$, $\bar{\mu}_N = N^{-1}\sum_{i\in\Omega}\mu_i$, $\bar{y}_N = N^{-1}\sum_{i\in\Omega} y_i$.
Variance (18) may be estimated by:

$$\hat{D}_p^2\left(\hat{t}^{MCAL}\right) = \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j>i}}^{n}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{u_i}{\pi_i} - \frac{u_j}{\pi_j}\right)^2, \tag{19}$$

where: $u_i = y_i - \hat{\mu}_i \hat{B}_N$.

Wu [2003] proves that the model-calibration estimator of the population total is optimal in the class of the calibration estimators in the sense that it minimizes model-expected asymptotic design-variance.

4. CALIBRATION AND MODEL-CALIBRATION ESTIMATORS OF DOMAIN TOTAL FOR DATA FROM ONE PERIOD

In the case of estimating domain totals there are at least three possible ways of using approaches presented in the previous sections.
The first one is:

$$\hat{t}_d^{GREG} = \sum_{i\in s_d} w_{si} y_i, \tag{20}$$

where w_{si} are weights of the calibration estimator given by (4), but used not for all of the sampled elements, but for elements sampled from the domain of interest, which gives the GREG estimator of the domain total. To estimate approximate design-variance of (20). Rao[9] suggests using estimator (10), where e_i should be replaced by $e_{id} = a_{id*} y_i - \mathbf{x}_i^T \hat{\mathbf{B}}'$,
where: $a_{id*} = \begin{cases} 1 & \text{for } i \in \Omega_{d*} \\ 0 & \text{for } i \notin \Omega_{d*} \end{cases}$, Ω_{d*} is the domain of interest, $\hat{\mathbf{B}}'$ is given by (7), where y_i are replaced by $a_{id*} y_i$. In the case of the model-calibration estimator of the population total it should be noted that even in the case of a linear model, but one having domain specific parameters, this estimator is a nonlinear function of y_i. Hence, an estimator defined similarly to (20) is used, but model-calibration is problematic.

[9] Rao J.N.K., [2003], p.17

The second proposal is to look for weights conditionally minimizing some distance measure between the estimator's weights and the sampling weights, however not for the whole sample (as in (3)), but for the sample in the domain, where the constraint is given by a calibration equation similar to (2), but defined for the domain. It may be written as follows:

$$\begin{cases} \sum\limits_{i\in s_d} \dfrac{\left(w_{sdi}-d_i\right)^2}{d_i q_i} \to \min \\ \sum\limits_{i\in s_d} w_{sdi}\mathbf{x}_i = \sum\limits_{i\in \Omega_d} \mathbf{x}_i \end{cases}. \tag{21}$$

Solving (21) provides the following weights[10]:

$$w_{sdi} = g_{sdi} d_i\,, \tag{22}$$

where: $\mathbf{t}_{dx} = \sum\limits_{i\in\Omega_d} \mathbf{x}_i$, $\hat{\mathbf{t}}_{dx}^{HT} = \sum\limits_{i\in s_d} d_i\,\mathbf{x}_i$ and

$$g_{sdi} = 1 + (\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}^{HT})^T \left(\sum_{i\in s_d} d_i q_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i q_i\,. \tag{23}$$

Based on weights (22), the following formula of the GREG estimator of the domain total is obtained:

$$\hat{t}_d^{GREG\#} = \sum_{i\in s_d} w_{sdi} y_i = \hat{t}_{dy}^{HT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}^{HT}\right)^T \hat{\mathbf{B}}_d\,, \tag{24}$$

where: $\hat{t}_{dy}^{HT} = \sum\limits_{i\in s_d} d_i\, y_i$ and

$$\hat{\mathbf{B}}_d = \left(\sum_{i\in s_d} d_i q_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i\in s_d} d_i q_i \mathbf{x}_i y_i\ . \tag{25}$$

To estimate approximate design variance of the estimator (24), Rao[11] proposes to use the estimator given by (10), where e_i should be replaced by $e_{id}^{\#} = a_{id*} y_i - a_{id*}\mathbf{x}_i^T \hat{\mathbf{B}}_d$. Similarly, the model-calibration estimator may be obtained by solving:

$$\begin{cases} \sum\limits_{i\in s_d} \dfrac{\left(w_{sdi}-d_i\right)^2}{d_i q_i} \to \min \\ N_d^{-1} \sum\limits_{i\in s_d} w_i = 1 \ \wedge \ \sum\limits_{i\in s_d} w_{si}\mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = \sum\limits_{i\in\Omega_d} \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \end{cases}, \tag{26}$$

but under linear model (even with domain specific parameters) (26) simplifies to (21), where the vector of ones is included.

[10] Ibidem, p.18.
[11] Ibidem.

The third proposal uses GREG (MGREG)

$$\hat{t}_d^{MGREG} = \hat{t}_{dy}^{HT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}^{HT}\right)^T \hat{\mathbf{B}} = \mathbf{t}_{\mathbf{dx}}^T \hat{\mathbf{B}} + \sum_{i \in s_d} \frac{e_i}{\pi_i}, \tag{27}$$

(where $\hat{\mathbf{B}}$ is given by (7) and e_i is given by (11)) that has the following benchmarking property:

$$\sum_{d=1}^{D} \hat{t}_d^{MGREG} = \hat{t}^{GREG}, \tag{28}$$

where $\hat{t}^{GREG}$ is given by (6). Särndal and Hidiroglou [1989] propose to improve (27) by modifying somewhat the error term $\sum_{i \in s_d} \frac{e_i}{\pi_i}$. But their modification suffers when the domain sample size is small, unlike the case of the estimator (27)[12].

To estimate *p*-variance of (27) Rao[13] proposes to use (10), where e_i are replaced by $a_{id*} e_i$. This variance estimator is valid even if the small area sample size is small, provided that the overall sample size is large. Similarly, we propose a formula of the modified model calibration estimator of the domain total:

$$\hat{t}_d^{MMCAL} = \hat{t}_{dy}^{HT} + \left(\sum_{i \in \Omega_d} \hat{\mu}_i - \sum_{i \in s_d} d_i \hat{\mu}_i \right) \hat{B}_N, \tag{29}$$

(where $\hat{B}_N$ is given by (17)) which has the following benchmarking property:

$\sum_{d=1}^{D} \hat{t}_d^{MMCAL} = \hat{t}^{MCAL}$,

where $\hat{t}^{MCAL}$ is given by (16). To estimate *p*-variance of (29) we will use (19), where u_i will be replaced by $a_{id*} u_i$.

5. MODEL FOR LONGITUDINAL DATA

In the paper, longitudinal data for periods t=1,...,M are considered. In the period t the population of size N_t is denoted by Ω_t. The population in the period t is divided into D disjoint domains (subpopulations) Ω_{dt} of size N_{dt}, where d=1,...,D. Let the set of population elements for which observations are available in the period t be denoted by s_t and its size by n_t. The set of the domain elements for which observations are available in the period t is denoted by s_{dt} and its size by n_{dt}.

We assume that the population may change in time and that one population element may change its domain membership in time (from a technical point of

[12] Ibidem, p.21.
[13] Ibidem.

view, observations of some population element which changes its domain membership are treated as observations of a new population element). This means that *i* and *t* completely identify domain membership, but additional subscript *d* will be needed as well. Let M_{id} denote the number of periods when the *i*-th population element may be potentially observed in the *d*-th domain (when the *i*-th population element belongs to the *d*-th domain). Let us denote the number of periods when the *i*-th population element (which belongs to the *d*-th domain) is observed by m_{id}. Let $m_{rid} = M_{id} - m_{id}$.

Values of the variable of interest are realizations of random variables Y_{idj} for the *i*-th population element which belongs to the *d*-th domain in the period t_{ij}, where $i=1,...,N$, $j=1,...,M_{id}$, $d=1,...,D$. The vector of size $M_{id} \times 1$ of random variables Y_{idj} for the *i*-th population element which belongs to the *d*-th domain will be denoted by $\mathbf{Y}_{id} = \left[Y_{idj} \right]$, where $j = 1,..., M_{id}$.

We consider superpopulation models used for longitudinal data[14], which are special cases of the GLM and the General Linear Mixed Model (GLMM). The following two-stage model is assumed. Firstly:

$$\mathbf{Y}_{id} = \mathbf{Z}_{id}\boldsymbol{\beta}_{id} + \mathbf{e}_{id}, \tag{30}$$

where $i=1,...,N$; $d=1,...,D$, $\mathbf{Y}_{id}$ is a random vector of size $M_{id} \times 1$, $\mathbf{Z}_{id}$ is known matrix of size $M_{id} \times q$, $\boldsymbol{\beta}_{id}$ is a vector of unknown parameters of size $q \times 1$, $\mathbf{e}_{id}$ is a random component vector of size $M_{id} \times 1$. Vectors $\mathbf{e}_{id}$ ($i=1,...,N$; $d=1,...,D$) are independent with $\mathbf{0}$ vector of expected values and variance-covariance matrix $\mathbf{R}_{id}$. Although $\mathbf{R}_{id}$ may depend on *i*, it is often assumed that $\mathbf{R}_{id} = \sigma_e^2 \mathbf{I}_{M_{id}}$ where $\mathbf{I}_{M_{id}}$ is the identity matrix of rank M_{id}. Secondly, we assume that:

$$\boldsymbol{\beta}_{id} = \mathbf{K}_{id}\boldsymbol{\beta} + \mathbf{v}_{id}, \tag{31}$$

where $i=1,...,N$; $d=1,...,D$, $\mathbf{K}_{id}$ is known matrix of size $q \times p$, $\boldsymbol{\beta}$ is a vector of unknown parameters of size $p \times 1$, $\mathbf{v}_{id}$ is a vector of random components of size $q \times 1$. It is assumed that vectors $\mathbf{v}_{id}$ ($i=1,...,N$; $d=1,...,D$) are independent with $\mathbf{0}$ vector of expected values and variance-covariance matrix $\mathbf{G}_{id} = \mathbf{H}$, which means that $\mathbf{G}_{id}$ does not depend on *i*.

[14] For example: Verbeke G., Molenberghs G., [2000], *Linear Mixed Models for Longitudinal Data*, Springer-Verlag, New York; Hedeker D., Gibbons R.D., [2006], *Longitudinal Data Analysis*, John Wiley & Sons, Hoboken, New Jersey.

Verbeke, Molenberghs[15] present similar assumptions to (30) and (31), however 3 differences exist. Firstly, in the book assumptions are made for profiles defined by elements. In this paper, assumptions are made for profiles defined by elements and domain membership, i.e. $\mathbf{Y}_{id}$ (of size $M_{id} \times 1$). Secondly, in the book the assumptions are made only for the sampled elements (i.e. i=1,...,n). In this paper they are made for all population elements (i=1,...,N). Thirdly, the notations by Verbeke and Molenberghs [2000] do not take into account (unlike this paper) the possibility of population changing in time.

Based on (30) and (31), we obtain:

$$\mathbf{Y}_{id} = \mathbf{X}_{id}\boldsymbol{\beta} + \mathbf{Z}_{id}\mathbf{v}_{id} + \mathbf{e}_{id}, \tag{32}$$

where i=1,...,N; d=1,...,D, $\mathbf{X}_{id} = \mathbf{Z}_{id}\mathbf{K}_{id}$ is known matrix of size $M_{id} \times p$. Let $\mathbf{V}_{id} = D^2_{\xi}(\mathbf{Y}_{id})$. Hence, $\mathbf{V}_{id} = \mathbf{Z}_{id}\mathbf{H}\mathbf{Z}^T_{id} + \mathbf{R}_{id}$.

Let $\mathbf{A}_d$ be a column vector and $col_{1\le d\le D}(\mathbf{A}_d) = \left[\mathbf{A}_1^T \ \ ... \ \ \mathbf{A}_d^T \ \ ... \ \ \mathbf{A}_D^T\right]^T$ be a column vector obtained by stacking $\mathbf{A}_d$ vectors. Note that by stacking $\mathbf{Y}_{id}$ vectors (i.e. $\mathbf{Y} = col_{1\le d\le D}(col_{1\le i\le N_d}(\mathbf{Y}_{id}))$ from (32) we obtain the formula of the GLMM.

6. CALIBRATION AND MODEL-CALIBRATION ESTIMATORS FOR LONGITUDINAL DATA

Under model (32), to estimate domain total in period t we may use:

(i) GREG direct estimator given by (20) (which will be denoted by GREGd),
(ii) GREG direct estimator given by (20), where the calibration equation includes auxiliary variables from all periods (which will be denoted by GREGd4 and omitted from the simulation study due to the small domain sample sizes),
(iii) GREG# indirect estimator given by (24) (which will be denoted by GREGi),
(iv) GREG# indirect estimator given by (24), where the calibration equation includes auxiliary variables not from one, but from all periods (which will be denoted by GREGi4),
(v) MGREG estimator given by (27) (which will be denoted by MGREGpop),
(vi) Modified MGREG estimator given by (27), but with $\hat{\mathbf{B}}$ not obtained using all sample information from the period of interest (as in (27)), but using sample information from the domain of interest from all periods according to the following formula:

[15] Verbeke G., Molenberghs G., [2000], p.20.

$$\hat{\mathbf{B}} = \left(\sum_{i \in \bigcup_{j=1}^{m} s_{d*j}} d_i q_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i \in \bigcup_{j=1}^{m} s_{d*j}} d_i q_i \mathbf{x}_i y_i \ , \tag{33}$$

(which will be denoted by MGREGdom),

(vii) Modified Model Calibrating estimator (29) (which will be denoted by MMCALpop), where:

– $\hat{B}_N$ given by (17) is estimated using y's from the period of interest,

– $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}$ is design-based estimator of $\boldsymbol{\beta}$ (see (32)) given by general formula (14), where information (on x's, y's and π_i's) from all periods is used (which will be denoted by MMCALpop)

(viii) Modified Model Calibrating estimator (29) (the estimator will be denoted by MMCALdom) where:

– $\hat{B}_N$ given by (17) is replaced by:

$$\hat{B}_N = \left(\sum_{i \in \bigcup_{j=1}^{m} s_{d*j}} d_i q_i (\hat{\mu}_i - \overline{\mu})^2 \right)^{-1} \sum_{i \in \bigcup_{j=1}^{m} s_{d*j}} d_i q_i (\hat{\mu}_i - \overline{\mu})(y_i - \overline{y}) \ ,$$

– $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}$ is design-based estimator of $\boldsymbol{\beta}$ (see (32)) given by general formula (14), where information (on x's, y's and π_i's) from all periods is used (which will be denoted by MMCALdom).

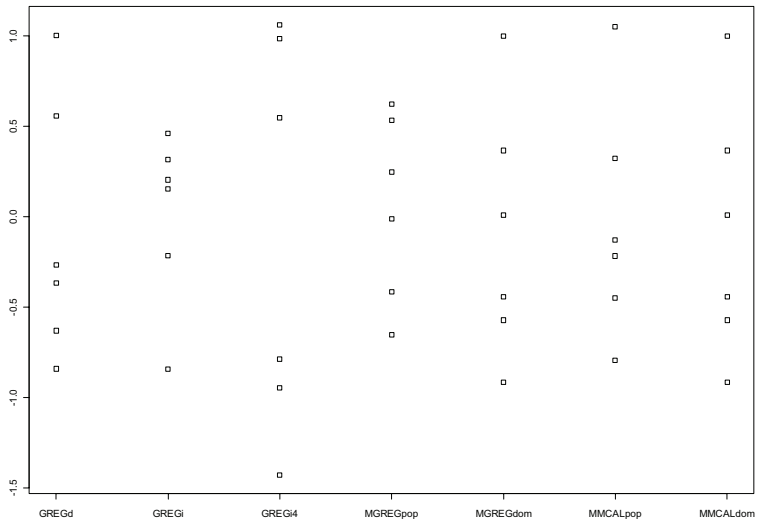
It is worth noticing that estimators GREGd4, GREGi4, MGREGdom, MMCALpop and MMCALdom are new proposals of calibration estimators for longitudinal data obtained by modifying known calibration or model calibration estimators.

In the case of estimators GREGd, GREGi and MGREGpop we use information on the variable of interest and auxiliary variables only from the period of interest. In the case of estimators GREGd4 and GREGi4 we use information on the variable of interest from the period of interest and auxiliary variables from all periods. In the case of estimators MGREGdom, MMCALpop and MMCALdom we use information on the variable of interest and auxiliary variables from all periods.

7. SIMULATION ANALYSIS

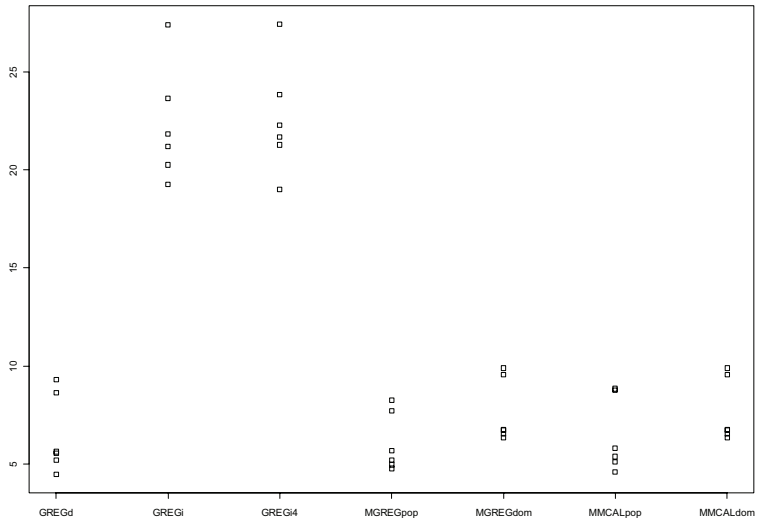
The Monte Carlo simulation analysis based on real data on N=314 Polish poviats (excluding cites with poviat's rights), which represent NTS 4 level, for M=4 years 2005-2008 (data derived from www.stat.gov.pl).

Graph 1. Relative design biases of estimators (%) of considered estimators in 6 domains



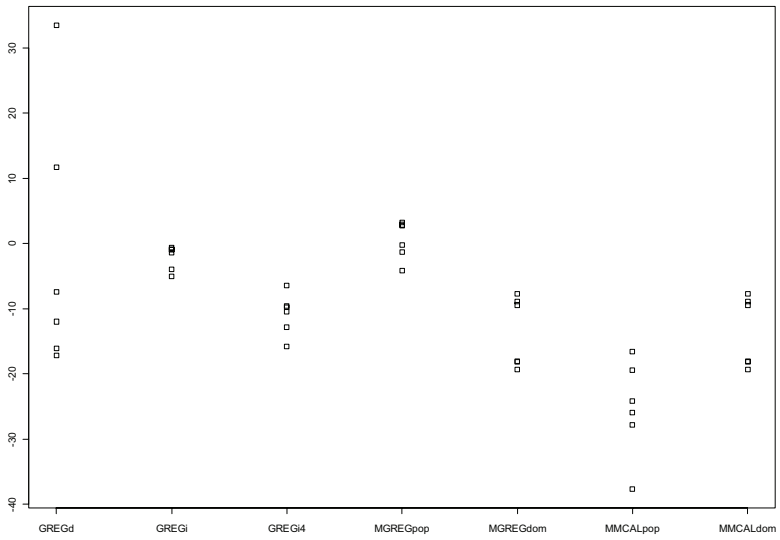
Source: developed by the author.

Graph 2. Relative design RMSE (%) of considered estimators in 6 domains



Source: developed by the author.

Graph 3. Relative design biases of MSE estimators (%) in 6 domains



Source: developed by the author.

The problem is to estimate subpopulations (domains) totals for D=6 regions (NTS 1 level) in 2008. The variable of interest is the poviats' own revenues (in PLN) and the auxiliary variable is the population size in the poviats (in persons). In the graphs below, each point represents the value of some statistics for one out of six domains. The simulation is design-based. In this case, a sample of the size n=79 elements (ca 25% of population size) is a balanced panel sample drawn at random in the first period with inclusion probabilities proportional to the value of the auxiliary variable in this period. With this sample size, it was possible to estimate all domain totals in each iteration, even using direct estimators. The number of samples drawn in the simulation equals 10 000.

In the simulation, absolute relative design biases of all estimators are smaller than 2%. Comparing the values of design relative RMSE we note that the highest accuracy is obtained for GREGd, MGREGpop and MMCALpop. When these 3 estimators are compared, then the absolute relative biases of design MSEs estimators are the smallest, on average, for MGREGpop. MGREGdom and MMCALdom use sample domain information only from 4 periods and may be more accurate for data with a larger number of periods.

8. SUMMARY

In the paper, several modification of calibration and model-calibration estimators of domain total for longitudinal data are proposed along with estimators of design MSE. Their accuracies are compared for real longitudinal data on Polish poviats.

REFERENCES

Deville J.C., Särndal C.E., [1992], *Calibration estimators in survey sampling*, Journal of the American Statistical Association, 87, pp.376-382.

Hedeker D., Gibbons R.D., [2006], *Longitudinal Data Analysis*, John Wiley & Sons, Hoboken, New Jersey.

Rao J.N.K., [2003], *Small area estimation*. John Wiley & Sons, New York.

Särndal C.E., Hidiroglou M.A., [1989], *Small domain estimation: A conditional analysis*, Journal of the American Statistical Association, 84, pp.266-275.

Särndal C.E., Swensson B., Wretman J., [1992], *Model assisted survey sampling*, Springer-Verlag, New York.

Singh A.C., Mohl C.A., [1996], *Understanding calibration estimators in survey sampling*, Survey methodology, 22, pp.107-115.

Stukel D.M., Hidiroglou M.A., Särndal C.E., [1996], *Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization*, Survey methodology, 22, pp.177-125.

Theberge A., [2000], *Calibration and restricted weights*, Survey methodology, 26, pp.99-107.

Verbeke G., Molenberghs G., [2000], *Linear Mixed Models for Longitudinal Data*, Springer-Verlag, New York.

Wu C., Sitter R.R., [2001], *A model-calibration approach to using complete auxiliary information from survey data*, Journal of the American Statistical Association, 96, pp.185-193.

Wu C., [2003], *Optimal calibration estimators in survey sampling*, Biometrika, 90, 4, pp.937-951.

O PEWNYCH ESTYMATORACH KALIBROWANYCH WARTOŚCI GLOBALNEJ W PODPOPULACJI W OPARCIU O DANE PRZEKROJOWO-CZASOWE

W artykule rozważane są modyfikacje znanych estymatorów kalibrowanych wartości globalnej w domenie na przypadek danych wielookresowych (w tym zmodyfikowany estymator kalibrowany modelowo). Dokładność zaproponowanych estymatorów została porównana z wykorzystaniem rzeczywistych danych wielookresowych. Najważniejsze rezultaty teoretyczne są prezentowane w części 3 (wzór opisujący zmodyfikowany estymator kalibrowany modelowo wartości globalnej w domenie i estymator jego *p*-błędu średniokwadratowego), w części 4 (zaproponowany model nadpopulacji) i w części 5 (nowe propozycje estymatorów dla danych wielookresowych).