Composite calibration estimation integrating data from different surveys

Takis Merkouris
Athens University of Economics and Business,
Patision 76, Athens 10434, Greece, email: merkouris@aueb.gr

Abstract

Current trends in survey sampling show a growing interest in integrating data from different surveys for improved estimation and analysis of population characteristics. For any case of different surveys sharing common items, we propose micro-integration of data through a suitable calibration scheme for the sampling weights of the combined sample, which produces a set of weights that incorporate all available information in the various surveys. These weights can be used to calculate weighted statistics, including totals, means, ratios, quantiles and regression coefficients. In particular, we obtain composite estimators of population totals that are asymptotically best linear unbiased estimators, or more practical composite estimators that are generalized regression estimators of a specific type and for certain sampling designs asymptotically best linear unbiased estimators. The construction of the calibration estimators is explained, and their statistical and computational efficiency is also discussed.

key words: Best linear unbiased estimation, combined samples, generalized regression estimation, matrix sampling, sampling weights.

1 Introduction

Integration of survey data may be generally defined as some combination of information from various surveys. Possibilities for such integration are on the increase in contemporary survey practice, primarily in social surveys. This is because there is a growing interest in integrating field work and survey functions of various survey sources for diverse reasons, such as, reduced cost and efficient survey operations, reduced response burden and improved data quality, harmonized survey content and data consistency, and improved estimation and analysis. A compilation of examples of data integration is presented in Merkouris (2010a).

This paper concerns integration of survey data for the improvement of estimation of population characteristics, particularly totals. Such improvement is possible when there are common items among the various surveys, by pooling data on these items and by exploiting the correlation between various items of these surveys. A statistically and computationally efficient micro-integration of data through an appropriate adjustment of the survey weights can be accomplished by a suitable calibration scheme, which is based on the principles of best linear unbiased estimation and generalized regression estimation.

Underlying the integration of survey data for improved estimation, is the assumption that combined data on common items are comparable, or, as usually called, harmonized. It is to be noted that harmonization of data is increasingly practiced in National Statistical Agencies.

In Section 2, a general calibration methodology for composite estimation and its connection to best linear unbiased estimation and generalized regression estimation is demonstrated through two paradigms of data integration. The efficiency of the proposed calibration estimators is discussed in Section 3. Concluding remarks are made in Section 4.

2 Best linear unbiased estimation and calibration

2.1 A basic example

Consider first a basic setting of data integration, involving two surveys with samples S_1 and S_2, not necessarily independent, drawn from the same population with arbitrary designs and sizes n_1 and n_2. A vector of variables $\mathbf{x}$ and a vector of variables $\mathbf{y}$ are surveyed in S_1 and S_2, respectively, and a vector of variables $\mathbf{z}$ is surveyed in both samples. We denote by $\mathbf{w}_i$ the vector of design weights for sample S_i, $i = 1, 2$, and by $\mathbf{X}_i$, $\mathbf{Y}_i$ and $\mathbf{Z}_i$ the sample matrices of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, the subscripts indicating the sample. We denote by $\hat{\mathbf{X}}$ the Horvitz-Thompson (HT) estimator $\mathbf{X}_1'\mathbf{w}_1$ of the total $\mathbf{t_x}$ of $\mathbf{x}$, by $\hat{\mathbf{Y}}$ the HT estimator of the total $\mathbf{t_y}$ of $\mathbf{y}$, and by $\hat{\mathbf{Z}}_i$ the two HT estimators (based on S_i) of the total $\mathbf{t_z}$ of $\mathbf{z}$. For more efficient estimation of the totals $\mathbf{t_x}$, $\mathbf{t_y}$ and $\mathbf{t_z}$ we seek composite estimators that combine all the available information on $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ in the two samples. Such composite estimators that are best linear unbiased estimators (BLUE), i.e., minimum-variance linear unbiased combinations of the four estimators $\hat{\mathbf{X}}$, $\hat{\mathbf{Z}}_1$, $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Z}}_2$, are denoted by $\hat{\mathbf{X}}^B$, $\hat{\mathbf{Y}}^B$ and $\hat{\mathbf{Z}}^B$ and given in matrix form by

$$(\hat{\mathbf{X}}^{B\prime}, \hat{\mathbf{Y}}^{B\prime}, \hat{\mathbf{Z}}^{B\prime})' = \boldsymbol{\mathcal{P}}(\hat{\mathbf{X}}', \hat{\mathbf{Z}}_1', \hat{\mathbf{Y}}', \hat{\mathbf{Z}}_2')' \tag{1}$$

where $\boldsymbol{\mathcal{P}} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1}$, the matrix $\mathbf{W}$ has entries 1's and 0's and satisfies $E[(\hat{\mathbf{X}}', \hat{\mathbf{Z}}_1', \hat{\mathbf{Y}}', \hat{\mathbf{Z}}_2')]' = \mathbf{W}(\mathbf{t}_\mathbf{x}', \mathbf{t}_\mathbf{y}', \mathbf{t}_\mathbf{z}')'$, and $\mathbf{V}$ is the covariance matrix of $(\hat{\mathbf{X}}', \hat{\mathbf{Z}}_1', \hat{\mathbf{Y}}', \hat{\mathbf{Z}}_2')'$. It follows that $\text{Var}[(\hat{\mathbf{X}}^{B\prime}, \hat{\mathbf{Y}}^{B\prime}, \hat{\mathbf{Z}}^{B\prime})'] = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}$. Such an approach to composite estimation has been explored in two specific contexts of survey sampling; see Wolter (1979), Jones (1980) and Fuller (1990), Chipperfield and Steel (2009) and Merkouris (2013).

A more practical formulation of this estimation procedure is as follows. Using the condition of unbiasness, $E(\hat{\mathbf{X}}^B) = \mathbf{t_x}$, $E(\hat{\mathbf{Y}}^B) = \mathbf{t_y}$ and $E(\hat{\mathbf{Z}}^B) = \mathbf{t_z}$, we obtain the matrix $\boldsymbol{\mathcal{P}}$ of the coefficients in the linear combinations in (1) as

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \mathbf{I} & \mathbf{B_x} & \mathbf{0} & -\mathbf{B_x} \\ \mathbf{0} & \mathbf{B_y} & \mathbf{I} & -\mathbf{B_y} \\ \mathbf{0} & \mathbf{B_z} & \mathbf{0} & \mathbf{I} - \mathbf{B_z} \end{pmatrix},$$

and (1) takes then the form

$$(\hat{\mathbf{X}}^{B\prime}, \hat{\mathbf{Y}}^{B\prime}, \hat{\mathbf{Z}}^{B\prime})' = (\hat{\mathbf{X}}', \hat{\mathbf{Y}}', \hat{\mathbf{Z}}_2')' + \boldsymbol{\mathcal{B}}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2), \tag{2}$$

where $\boldsymbol{\mathcal{B}}$ is the second column in $\boldsymbol{\mathcal{P}}$, with optimal (variance-minimizing) value $\boldsymbol{\mathcal{B}}^o = -\text{Cov}[(\hat{\mathbf{X}}', \hat{\mathbf{Y}}', \hat{\mathbf{Z}}_2')', \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2][\text{Var}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2)]^{-1}$. Thus, an explicit expression of (2) is

$$\begin{aligned} \hat{\mathbf{X}}^B &= \hat{\mathbf{X}} - \text{Cov}[(\hat{\mathbf{X}}, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2][\text{Var}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2)]^{-1}[\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2] \\ \hat{\mathbf{Y}}^B &= \hat{\mathbf{Y}} - \text{Cov}[(\hat{\mathbf{Y}}, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2][\text{Var}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2)]^{-1}[\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2] \\ \hat{\mathbf{Z}}^B &= \hat{\mathbf{Z}}_2 - \text{Cov}[(\hat{\mathbf{Z}}_2, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2][\text{Var}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2)]^{-1}[\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2], \end{aligned} \tag{3}$$

from which it follows that

$$\begin{array}{rcl} \text{Var}(\hat{\mathbf{X}}^B) & = & \text{Var}(\hat{\mathbf{X}}) - \text{Cov}[(\hat{\mathbf{X}}, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2][\text{Var}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2)]^{-1}\text{Cov}[(\hat{\mathbf{X}}, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2] \\ \text{Var}(\hat{\mathbf{Y}}^B) & = & \text{Var}(\hat{\mathbf{Y}}) - \text{Cov}[(\hat{\mathbf{Y}}, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2][\text{Var}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2)]^{-1}\text{Cov}[(\hat{\mathbf{Y}}, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2] \\ \text{Var}(\hat{\mathbf{Z}}^B) & = & \text{Var}(\hat{\mathbf{Z}}_2) - \text{Cov}[(\hat{\mathbf{Z}}_2, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2][\text{Var}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2)]^{-1}\text{Cov}[(\hat{\mathbf{Z}}_2, \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2]. \end{array} \quad (4)$$

It will give a hint for later development to notice that $\hat{\mathbf{Z}}^B$ can be alternatively written as $\hat{\mathbf{Z}}_1 - (\mathbf{I} - \mathbf{B}_{\mathbf{z}}^o)[\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_2] = \mathbf{B}_{\mathbf{z}}^o\hat{\mathbf{Z}}_1 + (\mathbf{I} - \mathbf{B}_{\mathbf{z}}^o)\hat{\mathbf{Z}}_2$, in obvious notation for $\mathbf{B}_{\mathbf{z}}^o$, and that substitution of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ by $\hat{\mathbf{Z}}$ in (3) gives $\hat{\mathbf{Z}}^B$. The expressions in (4) show the efficiency of $\hat{\mathbf{Z}}^B$ relative to $\hat{\mathbf{Z}}_1$ and $\hat{\mathbf{Z}}_2$, and the dependence of the efficiency of $\hat{\mathbf{X}}^B$ and $\hat{\mathbf{Y}}^B$, relative to $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, on the strength of the correlation of $\mathbf{z}$ with $\mathbf{x}$ and $\mathbf{y}$. The obvious modifications in (3) and (4) apply when the samples are independent.

It works out that the matrix $\boldsymbol{\mathcal{B}}^o$ can be written as $\boldsymbol{\mathcal{B}}^o = \text{Cov}(\hat{\boldsymbol{\mathcal{X}}}, \hat{\boldsymbol{\mathcal{Z}}})[\text{Var}(\hat{\boldsymbol{\mathcal{Z}}})]^{-1}$, where

$$\boldsymbol{\mathcal{X}} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}_2 & \mathbf{Z}_2 \end{pmatrix}, \qquad \boldsymbol{\mathcal{Z}} = \begin{pmatrix} -\mathbf{Z_1} \\ \mathbf{Z_2} \end{pmatrix},$$

and then an estimated optimal $\hat{\boldsymbol{\mathcal{B}}}^o$ takes the form $\hat{\boldsymbol{\mathcal{B}}}^o = (\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}})(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}})^{-1}$, where $\boldsymbol{\Lambda}^0 = \{(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l\pi_{kl}\}$, and π_k, π_{kl} are first-and-second order inclusion probabilities. The matrix $\boldsymbol{\Lambda}^0$ is associated with the combined sample $S = S_1 \cup S_2$, and for independent samples reduces to the block-diagonal matrix $\text{diag}\{\boldsymbol{\Lambda}_i^0\}$, with $\boldsymbol{\Lambda}_i^0$ associated with S_i. With this estimated $\hat{\boldsymbol{\mathcal{B}}}^o$, the right hand side of (2) is written as $\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}})^{-1}\hat{\boldsymbol{\mathcal{Z}}} = \boldsymbol{\mathcal{X}}'[\mathbf{w} + \boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{Z}}'\mathbf{w})]$, where $\mathbf{w}$ is the vector of design weights of the combined sample S. It appears that the estimated BLUE in (2) has the form of a calibration estimator, with vector of calibrated weights $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{Z}}'\mathbf{w})$ that implies a zero difference between the two estimates of the total $\mathbf{t}_{\mathbf{z}}$. Indeed, it can be shown that the vector $\mathbf{c}$ minimizes the generalized least-squares distance $(\mathbf{c} - \mathbf{w})'(\boldsymbol{\Lambda}^0)^{-1}(\mathbf{c} - \mathbf{w})$ while satisfying the constraint $\mathbf{Z}_1'\mathbf{c}_1 = \mathbf{Z}_2'\mathbf{c}_2$, where the subvector $\mathbf{c}_i$ corresponds to sample S_i. We may now write the estimated BLUE formally as a calibration estimator, $\hat{\boldsymbol{\mathcal{X}}}^B = \boldsymbol{\mathcal{X}}'\mathbf{c}$, and its three components as $\hat{\mathbf{X}}^B = \mathbf{X}_1'\mathbf{c}_1$, $\hat{\mathbf{Y}}^B = \mathbf{Y}_2'\mathbf{c}_2$ and $\hat{\mathbf{Z}}^B = \mathbf{Z}_1'\mathbf{c}_1 = \mathbf{Z}_2'\mathbf{c}_2$. Notice that the vector $\mathbf{c}$ can be written in the "residual" form $\mathbf{c} = (\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})'\mathbf{w}$, where $\mathbf{P}_{\boldsymbol{\mathcal{Z}}} = \boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{Z}})^{-1}\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}^0$, so that $\hat{\boldsymbol{\mathcal{X}}}^B = [(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})\boldsymbol{\mathcal{X}}]'\mathbf{w}$, and thus the estimated large sample variance of $\hat{\boldsymbol{\mathcal{X}}}^B$ is $\widehat{\text{Var}}(\hat{\boldsymbol{\mathcal{X}}}^B) = [(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})\boldsymbol{\mathcal{X}}]'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{Z}}})\boldsymbol{\mathcal{X}}$. Notably, by defining $\mathbf{w}$, $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\Lambda}^0$ at population level, the BLUE has the same form of calibration estimator as above.

The formulation of BLUE based on (2) has many advantages over the original form in (1). Explicit forms of estimates and variances show the difference between the BLUE,s and the simple HT estimators, as well as their relative efficiency. The computations of the estimators and their estimated variances is considerably easier. Importantly, any estimate is obtained in simple linear form as a weighted sum of the sample values of the corresponding variable, as in common practice in survey sampling. The established calibration property of the BLUE implies that estimates of totals for the common variables are consistent across surveys, often a desired property in the integration of survey data.

The two surveys may include vectors of auxiliary variables, $\mathbf{x}_a$ and $\mathbf{y}_a$ respectively, for which the vectors of population totals $\mathbf{t}_{\mathbf{x}_a}$ and $\mathbf{t}_{\mathbf{y}_a}$ are known. For improved efficiency, this auxiliary information is incorporated in the BLUE by including $\mathbf{t}_{\mathbf{x}_a} - \hat{\mathbf{X}}_a$ and $\mathbf{t}_{\mathbf{y}_a} - \hat{\mathbf{Y}}_a$ in the right hand side of (1), or (2). The resulting estimator is generated by a calibration

procedure that includes the constraints $\hat{\mathbf{X}}_a^B = \mathbf{t}_{\mathbf{x}_a}$ and $\hat{\mathbf{Y}}_a^B = \mathbf{t}_{\mathbf{y}_a}$, having the design matrix $\boldsymbol{\mathcal{Z}}$ augmented by the block-diagonal matrix $\boldsymbol{\mathcal{D}} = \text{diag}\{\mathbf{X}_a, \mathbf{Y}_a\}$, with corresponding vector of calibration totals $(\mathbf{0}^{'}, \mathbf{t}_{\boldsymbol{\mathcal{D}}}^{'})^{'}$, $\mathbf{t}_{\boldsymbol{\mathcal{D}}} = (\mathbf{t}_{\mathbf{x}_a}^{'}, \mathbf{t}_{\mathbf{y}_a}^{'})$. An equivalent form of the BLUE has the HT estimators in the right hand of (1) replaced by the optimal regression estimators $\hat{\mathbf{X}}^{OR}, \hat{\mathbf{Z}}_1^{OR}, \hat{\mathbf{Y}}^{OR}, \hat{\mathbf{Z}}_2^{OR}$, where, for example, $\hat{\mathbf{X}}^{OR} = \hat{\mathbf{X}} + \text{Cov}(\hat{\mathbf{X}}, \hat{\mathbf{X}}_a)[\text{Var}(\hat{\mathbf{X}}_a)]^{-1}[\mathbf{t}_{\mathbf{x}_a} - \hat{\mathbf{X}}_a]$ and $\hat{\mathbf{Z}}_1^{OR} = \hat{\mathbf{Z}}_1 + \text{Cov}(\hat{\mathbf{Z}}_1, \hat{\mathbf{X}}_a)[\text{Var}(\hat{\mathbf{X}}_a)]^{-1}[\mathbf{t}_{\mathbf{x}_a} - \hat{\mathbf{X}}_a]$; for a discussion of the optimal regression estimator, see Montanari (1987) and Rao (1994). The calibration form of the BLUE is then $\hat{\boldsymbol{\mathcal{X}}}^B = \boldsymbol{\mathcal{X}}^{'}\mathbf{c}$, where $\mathbf{c} = \mathbf{c}_a + \mathbf{L}^0\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}^{'}\mathbf{L}^0\boldsymbol{\mathcal{Z}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{Z}}^{'}\mathbf{c}_a]$ with $\mathbf{c}_a = \mathbf{w} + \boldsymbol{\Lambda}^0\boldsymbol{\mathcal{D}}(\boldsymbol{\mathcal{D}}^{'}\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{D}})^{-1}[(\mathbf{t}_{\boldsymbol{\mathcal{D}}}^{'} - \boldsymbol{\mathcal{D}}^{'}\mathbf{w}]$, and where $\mathbf{L}^0 = \boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{D}}})$ with $\mathbf{P}_{\boldsymbol{\mathcal{D}}} = \boldsymbol{\mathcal{D}}(\boldsymbol{\mathcal{D}}^{'}\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{D}})^{-1}\boldsymbol{\mathcal{D}}^{'}\boldsymbol{\Lambda}^0$.

Although the calibration procedure substantially facilitates the computation of the estimated BLUE of any total of interest, especially when the number of variables is large or when there are more samples involved, the matrix $\boldsymbol{\Lambda}^0$ makes the calculations quite laborious, particularly for dependent samples. Besides, the probabilities π_{kl} are not known for most complex sampling designs. Approximate forms of $\boldsymbol{\Lambda}^0$ that bypass this difficulty may well be used, provided that the approximate matrix is positive-(semi)definite. A computationally very convenient, but generally suboptimal, approach involves replacing $\boldsymbol{\Lambda}^0$ with the diagonal 'weighting matrix' $\boldsymbol{\Lambda}$ having w_{ik}/q_{ik} as ikth diagonal entry, where $\{w_{ik}\}$ are the design weights of S_i and $\{q_{ik}\}$ are positive constants. This gives the composite generalized regression (CGR) estimator of $(\mathbf{t}_{\mathbf{x}}^{'}, \mathbf{t}_{\mathbf{y}}^{'}, \mathbf{t}_{\mathbf{z}}^{'})^{'}$, which has the form of (2), but with the sample regression coefficient $\hat{\boldsymbol{\mathcal{B}}} = (\boldsymbol{\mathcal{X}}^{'}\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}})(\boldsymbol{\mathcal{Z}}^{'}\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}})^{-1}$ in place of $\hat{\boldsymbol{\mathcal{B}}}^o$. Note that $\hat{\boldsymbol{\mathcal{B}}}$ minimizes $(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}\boldsymbol{\mathcal{B}}^{'})^{'}\boldsymbol{\Lambda}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}\boldsymbol{\mathcal{B}}^{'})$, whereas $\hat{\boldsymbol{\mathcal{B}}}^o$ minimizes $(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}\boldsymbol{\mathcal{B}}^{'})^{'}\boldsymbol{\Lambda}^0(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{Z}}\boldsymbol{\mathcal{B}}^{'})$. In calibration form this estimator, $\hat{\boldsymbol{\mathcal{X}}}^{CGR} = \boldsymbol{\mathcal{X}}^{'}\mathbf{c}$, has calibration vector $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}^{'}\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{Z}}^{'}\mathbf{w})$. The value of q_{ik} in the entries of $\boldsymbol{\Lambda}_i$ should be set to $q_{ik} = \tilde{n}_i/(\tilde{n}_1 + \tilde{n}_2)$, where $\tilde{n}_i = n_i/d_i$, d_i denoting design effect, to take into account the differential in effective sample sizes between the two samples. If the same design is used for all samples, then $\tilde{n}_i = n_i$. The justification for this adjustment is based on an argument given in Merkouris (2010).

2.2 A matrix sampling example

In this setting of data integration, see Chipperfield and Steel (2009) and Merkouris (2013), all three vectors of variables $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are surveyed in an additional third sample S_3, and thus an additional HT estimator can be constructed for each of the totals $\mathbf{t}_{\mathbf{x}}$, $\mathbf{t}_{\mathbf{y}}$ and $\mathbf{t}_{\mathbf{z}}$. In analogy to (1), the BLUE in this case is

$$(\hat{\mathbf{X}}^{B\prime}, \hat{\mathbf{Y}}^{B\prime}, \hat{\mathbf{Z}}^{B\prime})^{'} = \boldsymbol{\mathcal{P}}(\hat{\mathbf{X}}_1^{'}, \hat{\mathbf{Z}}_1^{'}, \hat{\mathbf{Y}}_2^{'}, \hat{\mathbf{Z}}_2^{'}, \hat{\mathbf{X}}_3^{'}, \hat{\mathbf{Y}}_3^{'}, \hat{\mathbf{Z}}_3^{'})^{'},$$

and in line with the formulation of Section 2.1 it can be written in regression form as

$$(\hat{\mathbf{X}}^{B\prime}, \hat{\mathbf{Y}}^{B\prime}, \hat{\mathbf{Z}}^{B\prime})^{'} = (\hat{\mathbf{X}}_3^{'}, \hat{\mathbf{Y}}_3^{'}, \hat{\mathbf{Z}}_3^{'})^{'} + \boldsymbol{\mathcal{B}}[(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3)^{'}, (\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3)^{'}, (\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3)^{'}, (\hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3)^{'}]^{'}, \quad (5)$$

with optimal $\boldsymbol{\mathcal{B}}^o = -\text{Cov}(\mathbf{u}_3, \mathbf{u}_{12} - \mathbf{u}_3^{\star})[\text{V}(\mathbf{u}_{12} - \mathbf{u}_3^{\star})]^{-1}$, where $\mathbf{u}_3 = (\hat{\mathbf{X}}_3^{'}, \hat{\mathbf{Y}}_3^{'}, \hat{\mathbf{Z}}_3^{'})^{'}$, $\mathbf{u}_3^{\star} = (\hat{\mathbf{X}}_3^{'}, \hat{\mathbf{Z}}_3^{'}, \hat{\mathbf{Y}}_3^{'}, \hat{\mathbf{Z}}_3^{'})^{'}$, $\mathbf{u}_{12} = (\hat{\mathbf{X}}_1^{'}, \hat{\mathbf{Z}}_1^{'}, \hat{\mathbf{Y}}_2^{'}, \hat{\mathbf{Z}}_2^{'})^{'}$. It can be shown after some matrix algebra that the estimated $\boldsymbol{\mathcal{B}}^o$ is given by

$$\hat{\boldsymbol{\mathcal{B}}}^o = (\boldsymbol{\mathcal{X}}_{3-}^{'}\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{X}})(\boldsymbol{\mathcal{X}}^{'}\boldsymbol{\Lambda}^0\boldsymbol{\mathcal{X}})^{-1},$$

where

$$\boldsymbol{\mathcal{X}} = \begin{pmatrix} -\mathbf{X}_1 & -\mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{Y}_2 & -\mathbf{Z}_2 \\ \mathbf{X}_3 & \mathbf{Z}_3 & \mathbf{Y}_3 & \mathbf{Z}_3 \end{pmatrix},$$

is the design matrix for the estimator (5), $\boldsymbol{\mathcal{X}}_{3\_}$ is the matrix $\boldsymbol{\mathcal{X}}$ with the second column eliminated and the first two rows set equal to zero, and $\mathbf{\Lambda}^0$ is now associated with the combined sample $S = S_1 \cup S_2 \cup S_3$. The estimator (5) can be conveniently obtained through a calibration procedure that generates a vector of calibrated weights for the combined sample S, given by $\mathbf{c} = \mathbf{w} + \mathbf{\Lambda}^0\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{\Lambda}^0\boldsymbol{\mathcal{X}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}'\mathbf{w})$, and satisfying the constraints $\mathbf{X}_1'\mathbf{c}_1 = \mathbf{X}_3'\mathbf{c}_3$, $\mathbf{Y}_2'\mathbf{c}_2 = \mathbf{Y}_3'\mathbf{c}_3$ and $\mathbf{Z}_1'\mathbf{c}_1 = \mathbf{Z}_2'\mathbf{c}_2 = \mathbf{Z}_3'\mathbf{c}_3$. Expression (5) is then obtained simply as $\boldsymbol{\mathcal{X}}_{3\_}'\mathbf{c}$, based only on sample S_3. Replacing the matrix $\mathbf{\Lambda}^0$ in $\hat{\boldsymbol{\mathcal{B}}}$ and $\mathbf{c}$ with the weighting matrix $\mathbf{\Lambda}$, gives the CGR estimator and its calibration equivalent, respectively.

3 The efficiency of the calibration estimators

The calibration estimators of the previous section, the construction of which requires knowledge of $\mathbf{\Lambda}^0$, are asymptotically BLUE. For their linear form, they make the most efficient use of the available information in the integrated data.

The multivariate composite estimation for all components of the common vector $\mathbf{z}$, through the simultaneous calibration that aligns estimates from different surveys, may realize additional efficiency due to the correlations among the components of $\mathbf{z}$. This is unlike the single-sample setting, where estimation for any single variable may borrow strength only from auxiliary variables with known totals. Thus, in the context of Section 2.1, let $\hat{Z}_{i1}$ and $\hat{\mathbf{Z}}_{i^*}$ denote the HT estimators for the first component and for the remaining components of $\mathbf{z}$, respectively, based on sample S_i. When the two samples are independent, it can be shown (proof omitted) that the composite estimator $\hat{Z}_{11}^B$ (equal to $\hat{Z}_{21}^B$ by construction) is less efficient if only the first component of $\mathbf{z}$ is involved in the composition, unless $\text{Cov}(\hat{Z}_{11}, \hat{\mathbf{Z}}_{1^*})/\text{Var}(\hat{Z}_{11}) = \text{Cov}(\hat{Z}_{21}, \hat{\mathbf{Z}}_{2^*})/\text{Var}(\hat{Z}_{21})$. Since the quantities on each side of this equation are regression coefficients, the equation holds only if the effect of the regression of the first component of $\mathbf{z}$ on the rest is identical in samples S_1 and S_2. This can happen only if the sampling designs for the two samples are identical, including equal sample sizes, or only if the sampling design across samples is the same design with equal inclusion probability for all units, but not necessarily with the same sample size. The equation will not hold if the composite estimation incorporates auxiliary variables with known totals that are not identical in the two samples, owing to the associated differential regression effect. This effect will be accounted for by the proposed composite estimation method. Essentially the same property manifests itself in the composite estimation of section 2.2. Remarkably, the composite estimator $\hat{\mathbf{Z}}^B$ may then realize additional efficiency due to correlation of $\mathbf{z}$ with $\mathbf{x}$ and $\mathbf{y}$, although it incorporates all information on $\mathbf{z}$ available in the three samples. Indeed, it can be shown (Merkouris 2013) that the coefficients of the terms $\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3$ and $\hat{\mathbf{Y}}_1 - \hat{\mathbf{Y}}_3$ in (5) are zero only if $[\text{Var}(\hat{\mathbf{Z}}_1)]^{-1}\text{Cov}(\hat{\mathbf{X}}_1, \hat{\mathbf{Z}}_1) = [\text{Var}(\hat{\mathbf{Z}}_3)]^{-1}\text{Cov}(\hat{\mathbf{X}}_3, \hat{\mathbf{Z}}_3)$ and $[\text{Var}(\hat{\mathbf{Z}}_2)]^{-1}\text{Cov}(\hat{\mathbf{Y}}_2, \hat{\mathbf{Z}}_2) = [\text{Var}(\hat{\mathbf{Z}}_3)]^{-1}\text{Cov}(\hat{\mathbf{Y}}_3, \hat{\mathbf{Z}}_3)$.

In general, the computationally simpler CGR estimator, involving the coefficient $\hat{\boldsymbol{\mathcal{B}}}$, is less efficient than the estimated BLUE. On the other hand, the estimated BLUE may be unstable in small samples, when there is a small number of degrees of freedom available for the estimation of $\hat{\boldsymbol{\mathcal{B}}}^o$. This is akin to the issue of relative stability of the optimal versus the generalized regression estimator in the single-sample case; see Rao (1994) and Montanari (1998). For certain sampling strategies, $\hat{\boldsymbol{\mathcal{B}}} = \hat{\boldsymbol{\mathcal{B}}}^o$ and the CGR estimator is then the estimated BLUE. Proof of this property is offered in Merkouris (2013) for stratified (and unstratified) simple random sampling and Poisson sampling, with appropriate values in each case for the constants q_{ik} in the entries of $\mathbf{\Lambda}$.

4 Conclusion

The described estimation method for integrating information from different surveys involves a single-step calibration of the weights of the combined sample. Thus, using a single set of calibrated weights that incorporate all the available information from all samples, an improved estimate of the total for any variable, common or uncommon to the various surveys, can be obtained by using the units of only one of the samples containing the particular variable. These weights could be used to calculate other weighted statistics, including means, ratios, quantiles and regression coefficients. Composite estimates for domains of interest may be readily obtained by summing the weighted values of a variable over any of these domains. A simple modification of the calibration procedure that leads to more efficient composite estimation for domain totals of interest involves the augmentation of the design matrix with columns for the relevant variables defined at the domain level.

Estimation of the variance of an estimated BLUE, or of an CGR estimator, may, in principle, be based on the Taylor linearization technique. This approach requires calculations that are often not practical, or even feasible when the joint inclusion probabilities are not known. Replication methods for variance estimation, such as the jackknife method or the bootstrap method, could then be employed. For example, the jackknife method, customarily used in surveys with stratified multistage sampling design, can be adapted to replicate the calibration procedures that give rise to these composite estimators. This is fairly simple when the combined samples are independent, but rather complicated otherwise.

The proposed calibration approach may be readily extended to more complex settings of data integration, making more evident the operational power of the calibration procedure; the crucial step is to determine the design matrix. It may also encompass other, more traditional, settings of data integration, such as multiple-frame and two-phase sampling designs.

References

Chipperfield, J.O, and Steel, D.G. (2009) "Design and estimation for split questionnaire surveys," *Journal of Official Statistics*, 25, 227-244.

Fuller, W.A. (1990) "Analysis of repeated surveys," *Survey Methodology*, 16, 167-180.

Jones, R.G. (1980) "Best linear unbiased estimators for repeated surveys," *Journal of the Royal Statistical Society, Ser. B*, 42, 221-226.

Merkouris, T. (2010a) "Use of auxiliary sources in weighting: the case of integration of social survey data," In *Proceedings of Statistics Canada Symposium 2010. Social Statistics: The interplay among Censuses, Surveys and Administrative Data*, 256-263.

Merkouris, T. (2010b) "Combining information from multiple surveys by using regression for more efficient small domain estimation," *Journal of the Royal Statistical Society, Ser. B*, 72, 7-48.

Merkouris, T. (2013) "An efficient estimation method for matrix survey sampling," Submitted.

Montanari, G. E. (1987) "Post-sampling efficient QR-prediction in large-scale surveys," *International Statistics Review*, 55, 191-202.

Montanari, G. E. (1998) "On regression estimation of finite population means," *Survey Methodology*, 24, 69-77.

Rao, J. N. K. (1994) "Estimating totals and distribution functions using auxiliary information at the estimation stage," *Journal of Official Statistics*, 10, 153-165.

Wolter, K.M. (1979) "Composite estimation in finite populations," *Journal of the American Statistical Association*, 74, 604-613.