# Efficient Small-Domain Estimation by Combining Information from Multiple Surveys through Regression

Takis Merkouris
Statistics Canada

## Abstract

In sample surveys of finite populations, subpopulations for which the sample size is too small for estimation of adequate precision are referred to as small domains. In this paper, we explore the possibility of enhancing the precision of domain estimators by combining comparable information collected in multiple surveys of the same population. To this end, we propose a regression method of estimation that is essentially an extended calibration procedure whereby comparable domain estimates from the various surveys are calibrated to each other. We show through some analytic results that this method may greatly improve the precision of domain estimators for the variables that are common to these surveys, as these estimators make effective use of increased sample size for the common survey items. The proposed design-based direct estimators involve only domain-specific data on the variables of interest. The proposed approach is also highly effective in handling the closely related problem of estimation for rare population characteristics.

**Keywords**: Auxiliary information; Small area; Rare characteristics; Composite estimator; Generalized regression estimator; Calibration.

## 1 Introduction

National statistical agencies and other survey organizations regularly produce estimates for a number of subpopulations, called domains, as part of the statistical output of large scale surveys. Domain estimates are less precise than estimates for the whole population, primarily because of the smaller size of the associated sample and to a lesser degree because of the extra variability induced by the randomness of this sample size when the domains are not strata. For a particular survey, this shortcoming may limit the scope of domain estimation to rather large domains. On the other hand, demand for small-domain estimates has been growing in recent years among users of survey data. Since interest in small-domain estimation has traditionally centred on estimates for small geographic areas, the subject is generically referred to in the literature as small-area estimation.

Increasing demand for reliable small-area estimates has led over the past few years to the production of a sizeable literature on small-area estimation methods; a comprehensive account of such methods is given in Rao (2003). Invariably, these methods employ models to "borrow strength" for the variables of interest through the use of related survey data or administrative data that are external to the small areas of interest or are from other time periods. The derived domain estimators are then *indirect* estimators, in the sense that they incorporate data on the variables of interest that are external to the targeted small areas.

In this paper, we exploit the possibility of borrowing strength from other surveys of the same population that have collected comparable information on some or all variables of interest in the same domains. The potential for efficient small area estimation by combining comparable ("harmonized") information from multiple surveys has been recognized in recent literature (Marker 2001, Rao 2003, p.23) but there seems to be a paucity of related research up to now.

Combining information from multiple surveys for more precise estimation of survey characteristics at the population level has been the subject of recent research by Zieschang (1990), Renssen and Nieuwenbroek (1997) and Merkouris (2004), who used variants of generalized regression, and by Wu (2004) who took an empirical likelihood approach. In this paper, the regression procedure of Merkouris (2004) is adapted to small-domain estimation. The proposed regression method is essentially an extended calibration procedure whereby comparable domain estimates from the various surveys are calibrated to each other. Unlike the existing approaches to small-area estimation, borrowing strength with this method is not model-dependent, and the resulting domain estimators are direct as they involve only domain-specific data on the target variables. In particular, this design-based approach greatly enhances the reliability of domain estimators for the variables that are common to these surveys, as these estimators are based on increased effective sample size for the common survey items. The proposed estimation method is equally suitable for small geographic and non-geographic domains. It is also especially useful when dealing with the closely related problem of estimating rare population characteristics.

The organization of the paper is as follows. The notation and terminology are set out in Section 2. In Section 3, three variants of an extended GREG procedure are used to combine information from two surveys. The relative efficiency of these procedures is assessed analytically under certain conditions. In section 4, following a summary of the main findings, theoretical and practical aspects of the proposed estimation method are discussed.

## 2 Basic notation and terminology

Consider a finite population $U = \{1, \ldots, k, \ldots N\}$, from which a probability sample $s$ of size $n$ is drawn according to a sampling design with known first - and second - order inclusion probabilities $\pi_k$ and $\pi_{kl}$ $(k, l \in U)$. Consider the sampling weight vector $\mathbf{w}$ with $k$-th entry defined as $w_k = (1/\pi_k)I(k \in s)$, where $I$ denotes the indicator variable, and let $\mathbf{Y} \in \mathbb{R}^{N \times d}$ denote the population matrix of a $d$-dimensional survey variable of interest $\mathbf{y}$. The Horvitz - Thompson (HT) estimator of the total $\mathbf{t_y} = \mathbf{Y}'\mathbf{1}$, where $\mathbf{1}$ is the unit $N$-vector, is given by $\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{w}$ $(= \sum_U w_k y_k)$. For the population matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ of a $p$-dimensional auxiliary variable $\mathbf{x}$, assume that the total $\mathbf{t_x} = \mathbf{X}'\mathbf{1}$ is known. Let also $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ be the diagonal "weighting" matrix that has $w_k/q_k$ as $kk$-th entry, where $q_k$ is a positive constant, and use $s$ to designate the subvectors and submatrices corresponding to the sample. A vector of "calibrated" weights, $\mathbf{c}_s \in \mathbb{R}^n$, can be constructed to satisfy the constraints $\mathbf{X}_s'\mathbf{c}_s = \mathbf{t_x}$ while minimizing the generalized least squares distance $(\mathbf{c}_s - \mathbf{w}_s)'\mathbf{\Lambda}_s^{-1}(\mathbf{c}_s - \mathbf{w}_s)$. Assuming that $\mathbf{X}_s$ is of full rank $p$, this calibration procedure generates the vector

$$\mathbf{c}_s = \mathbf{w}_s + \mathbf{\Lambda}_s\mathbf{X}_s(\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}(\mathbf{t_x} - \mathbf{X}_s'\mathbf{w}_s). \qquad (1)$$

The calibration estimator of the total $\mathbf{t_y}$ is obtained as

$$\mathbf{Y}_s'\mathbf{c}_s = \mathbf{Y}_s'\mathbf{w}_s + \mathbf{Y}_s'\mathbf{\Lambda}_s\mathbf{X}_s(\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}(\mathbf{t_x} - \mathbf{X}_s'\mathbf{w}_s), \qquad (2)$$

which can take the form of a generalized regression (GREG) estimator

$$\hat{\mathbf{Y}}^R = \hat{\mathbf{Y}} + \hat{\boldsymbol{\beta}}(\mathbf{t_x} - \hat{\mathbf{X}}) = \hat{\boldsymbol{\beta}}\mathbf{t_x} + (\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}')'\mathbf{w}_s, \qquad (3)$$

where $\hat{\mathbf{X}} = \mathbf{X}_s'\mathbf{w}_s$ is the HT estimator of $\mathbf{t_x}$, and $\hat{\boldsymbol{\beta}} = \mathbf{Y}_s'\mathbf{\Lambda}_s\mathbf{X}_s(\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}$ is the matrix of sample regression coefficients. The term $(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}')'\mathbf{w}_s$ in (3) is the sum of weighted sample regression residuals. By construction the GREG estimator (3) has the calibration property that $\hat{\mathbf{X}}^R = \mathbf{t_x}$, that is, the GREG estimator of the total for $\mathbf{x}$ is equal to the known associated population total ("control" total). A formulation of the GREG estimator as a calibration estimator is given in Deville and Särndal (1992), and an extensive discussion of it is given in Särndal, Swensson, and Wretman (1992).

We define a domain $U_d$ to be any subset of $U$ and denote by $U_{\bar{d}}$ the complement of $U_d$. We let $\mathbf{Y}_d$ denote the matrix $\mathbf{Y}$ when for the $k$th row $\mathbf{y}_k = 0$ if $k \notin U_d$; accordingly, $\mathbf{Y}_{\bar{d}}$ denotes the matrix orthogonal to $\mathbf{Y}_d$. We can write then $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{Y}_d + \mathbf{Y}_{\bar{d}}$. Similarly for the matrix $\mathbf{X}$. Assuming that membership in $U_d$ for every sample unit is observed, we denote by $\mathbf{Y}_{sd}$ and $\mathbf{X}_{sd}$ the associated sample domain quantities. The HT estimator of the domain total $\mathbf{t}_{\mathbf{y}_d} = \mathbf{Y}_d\mathbf{1}$ is $\hat{\mathbf{Y}}_d = \mathbf{Y}_{sd}'\mathbf{w}_s$.

## 3 Domain estimation using information from two surveys

We assume that there exist samples $s_1$ and $s_2$ of sizes $n_1$ and $n_2$, respectively, from two independent surveys of the same target population, a vector $\mathbf{z}$ of $q$ survey variables common to $s_1$ and $s_2$, and auxiliary vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ associated with $s_1$ and $s_2$, respectively. Then, adapting the general procedure in Merkouris (2004) to estimation of totals in the domain $U_d$, we may combine information on $\mathbf{z}$ from the two samples using special regression setups for the combined sample as follows.

### 3.1 Regressing at the $U$ level and combining information at the $U_d$ level

A simultaneous regression for the two samples using the setup $\mathbf{X}_s = diag(\mathbf{X}_{s_i})$, $\mathbf{\Lambda}_s = diag(\mathbf{\Lambda}_{s_i})$, $\mathbf{w}_s = (\mathbf{w}_{s_1}', \mathbf{w}_{s_2}')'$, $\mathbf{t} = (\mathbf{t}_{\mathbf{x}_1}', \mathbf{t}_{\mathbf{x}_2}')'$, generates a vector of calibrated weights, $\mathbf{c}_{xs}$, given by

$$\mathbf{c}_{xs} = \begin{pmatrix} \mathbf{w}_{s_1} \\ \mathbf{w}_{s_2} \end{pmatrix} + \begin{pmatrix} \mathbf{\Lambda}_{s_1}\mathbf{X}_{s_1}(\mathbf{X}_{s_1}'\mathbf{\Lambda}_{s_1}\mathbf{X}_{s_1})^{-1}[\mathbf{t}_{\mathbf{x}_1} - \mathbf{X}_{s_1}'\mathbf{w}_{s_1}] \\ \mathbf{\Lambda}_{s_2}\mathbf{X}_{s_2}(\mathbf{X}_{s_2}'\mathbf{\Lambda}_{s_2}\mathbf{X}_{s_2})^{-1}[\mathbf{t}_{\mathbf{x}_2} - \mathbf{X}_{s_2}'\mathbf{w}_{s_2}] \end{pmatrix}. \qquad (4)$$

For any domain $U_d$, the two components of $\mathbf{c}_{xs}$ give rise to two independent GREG domain estimators $\hat{\mathbf{Z}}_{id}^R = \hat{\mathbf{Z}}_{id} + \mathbf{Z}_{s_{id}}'\mathbf{\Lambda}_{s_i}\mathbf{X}_{s_i}(\mathbf{X}_{s_i}'\mathbf{\Lambda}_{s_i}\mathbf{X}_{s_i})^{-1}[\mathbf{t}_{\mathbf{x}_i} - \hat{\mathbf{X}}_i]$, $i = 1, 2$, of the domain total $\mathbf{t}_{\mathbf{z}_d}$. Combining information on $\mathbf{z}$ at the domain level is accomplished by incorporating into the regression procedure the additional calibration constraint that the two estimators of $\mathbf{t}_{\mathbf{z}_d}$ are calibrated to each other, that is, they are aligned. This involves the extended regression matrix and the corresponding vector of control totals

$$\boldsymbol{\mathcal{X}}_s = \begin{pmatrix} \mathbf{X}_{s_1} & \mathbf{0} & \mathbf{Z}_{s_{1d}} \\ \mathbf{0} & \mathbf{X}_{s_2} & -\mathbf{Z}_{s_{2d}} \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1} \\ \mathbf{t}_{\mathbf{x}_2} \\ \mathbf{0} \end{pmatrix}. \qquad (5)$$

Now assume that $(\mathbf{X}_{s_i} \;\; \mathbf{Z}_{s_{id}})$ is of full rank $p_i + q$ and write $\boldsymbol{\mathcal{X}}_s$ in partition form as $\boldsymbol{\mathcal{X}}_s = (\mathbf{X}_s \;\; \boldsymbol{\mathcal{Z}}_{s_d})$, where $\mathbf{X}_s$ and $\boldsymbol{\mathcal{Z}}_{s_d}$ are of dimension $(n_1 + n_2) \times (p_1 + p_2)$ and $(n_1 + n_2) \times q$, respectively. Next let $\mathbf{L}_s = \mathbf{\Lambda}_s(\mathbf{I} - \mathbf{P}_{\mathbf{X}_s})$, with $\mathbf{P}_{\mathbf{X}_s} = \mathbf{X}_s(\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{\Lambda}_s$, and note that $\mathbf{X}_s = diag(\mathbf{X}_{s_i})$ implies $\mathbf{L}_s = diag(\mathbf{L}_{s_i})$, where $\mathbf{L}_{s_i} = \mathbf{\Lambda}_{s_i}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{s_i}})$, in obvious notation for $\mathbf{\Lambda}_{s_i}$ and $\mathbf{P}_{\mathbf{X}_{s_i}}$. Then, following Merkouris (2004), for weight vector $\mathbf{w}_s = (\mathbf{w}_{s_1}', \mathbf{w}_{s_2}')'$ and weighting matrix $\mathbf{\Lambda}_s = diag(\mathbf{\Lambda}_{s_i})$, the regression procedure based on the partitioned matrix $\boldsymbol{\mathcal{X}}_s$ generates the vector of calibrated weights

$$\begin{aligned} \mathbf{c}_s &= \mathbf{c}_{xs} + \mathbf{L}_s\boldsymbol{\mathcal{Z}}_{s_d}(\boldsymbol{\mathcal{Z}}_{s_d}'\mathbf{L}_s\boldsymbol{\mathcal{Z}}_{s_d})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{Z}}_{s_d}'\mathbf{c}_{xs}) \\ &= \begin{pmatrix} \mathbf{c}_{xs_1} \\ \mathbf{c}_{xs_2} \end{pmatrix} + \begin{pmatrix} \mathbf{L}_{s_1}\mathbf{Z}_{s_{1d}} \\ -\mathbf{L}_{s_2}\mathbf{Z}_{s_{2d}} \end{pmatrix} \left[\mathbf{Z}_{s_{1d}}'\mathbf{L}_{s_1}\mathbf{Z}_{s_{1d}} + \mathbf{Z}_{s_{2d}}'\mathbf{L}_{s_2}\mathbf{Z}_{s_{2d}}\right]^{-1} \\ &\qquad \times \left[(\mathbf{Z}_{s_{2d}}'\mathbf{c}_{xs_2} - \mathbf{Z}_{s_{1d}}'\mathbf{c}_{xs_1})\right]. \end{aligned}$$

It is easy to verify that the vector $\mathbf{c}_s$ satisfies all the calibration constraints, namely, $\mathbf{X}_{s_i}'\mathbf{c}_{s_i} = \mathbf{t}_{\mathbf{x}_i}$ and $\boldsymbol{\mathcal{Z}}_{s_d}'\mathbf{c}_s =$

$\mathbf{Z}'_{s_{1d}}\mathbf{c}_{s_1} - \mathbf{Z}'_{s_{2d}}\mathbf{c}_{s_2} = \mathbf{0}$. For any noncommon single variable $y_i$ associated with sample $s_i$, we can obtain composite GREG domain estimators $\hat{Y}^{CR}_{id} = \mathbf{Y}'_{s_{id}}\mathbf{c}_{s_i}$ of $\mathbf{t}_{y_{id}}$ that have the form

$$\begin{aligned} \hat{Y}^{CR}_{1d} &= \hat{Y}^{R}_{1d} + \hat{\mathbf{B}}_{y_{1d}}(\mathbf{I} - \hat{\mathbf{B}}_d)[\hat{\mathbf{Z}}^R_{2d} - \hat{\mathbf{Z}}^R_{1d}], \\ \hat{Y}^{CR}_{2d} &= \hat{Y}^{R}_{2d} - \hat{\mathbf{B}}_{y_{2d}}\hat{\mathbf{B}}_d[\hat{\mathbf{Z}}^R_{2d} - \hat{\mathbf{Z}}^R_{1d}], \end{aligned} \qquad (6)$$

where $\hat{\mathbf{B}}_{y_{id}} = \mathbf{Y}'_{s_{id}}\mathbf{L}_{s_i}\mathbf{Z}_{s_{id}}[\mathbf{Z}'_{s_{id}}\mathbf{L}_{s_i}\mathbf{Z}_{s_{id}}]^{-1}$, $\hat{\mathbf{B}}_d = \mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_2}\mathbf{Z}_{s_{2d}}[\mathbf{Z}'_{s_{1d}}\mathbf{L}_{s_1}\mathbf{Z}_{s_{1d}} + \mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_2}\mathbf{Z}_{s_{2d}}]^{-1}$ and $\hat{Y}^R_{id} = \hat{Y}_{id} + \mathbf{Y}'_{s_{id}}\mathbf{\Lambda}_{s_i}\mathbf{X}_{s_i}(\mathbf{X}'_{s_i}\mathbf{\Lambda}_{s_i}\mathbf{X}_{s_i})^{-1}[\mathbf{t}_{\mathbf{x}_i} - \hat{\mathbf{X}}_i]$. For the $q$-dimensional common variable $\mathbf{z}$ we have the two identical estimators of $\mathbf{t}_{\mathbf{z}_d}$

$$\begin{aligned} \hat{\mathbf{Z}}^{CR}_{1d} &= \hat{\mathbf{Z}}^{R}_{1d} + (\mathbf{I} - \hat{\mathbf{B}}_d)[\hat{\mathbf{Z}}^R_{2d} - \hat{\mathbf{Z}}^R_{1d}], \\ \hat{\mathbf{Z}}^{CR}_{2d} &= \hat{\mathbf{Z}}^{R}_{2d} - \hat{\mathbf{B}}_d[\hat{\mathbf{Z}}^R_{2d} - \hat{\mathbf{Z}}^R_{1d}], \end{aligned} \qquad (7)$$

which can be written in the form of the composite estimator

$$\hat{\mathbf{Z}}^{CR}_{1d} = \hat{\mathbf{Z}}^{CR}_{2d} = \hat{\mathbf{B}}_d\hat{\mathbf{Z}}^R_{1d} + (\mathbf{I} - \hat{\mathbf{B}}_d)\hat{\mathbf{Z}}^R_{2d}. \qquad (8)$$

The approximate design variance of $\hat{Y}^{CR}_{1d}$, denoted by $AV(\hat{Y}^{CR}_{1d})$, is given by

$$\begin{aligned} AV(\hat{Y}^{CR}_{1d}) = {} & AV(\hat{Y}^R_{1d}) \\ & + \mathbf{B}_{y_{1d}}(\mathbf{I} - \mathbf{B}_d)[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})] \\ & \times(\mathbf{I} - \mathbf{B}_d)'\mathbf{B}'_{y_{1d}} - 2\mathbf{B}_{y_{1d}}(\mathbf{I} - \mathbf{B}_d) \\ & \times[AC(\hat{Y}^R_{1d}, \hat{\mathbf{Z}}^R_{1d})]', \end{aligned} \qquad (9)$$

where $AC$ denotes approximate covariance and where $\mathbf{B}_{y_{1d}} = \mathbf{Y}'_d\mathbf{L}_1\mathbf{Z}_d[\mathbf{Z}'_d\mathbf{L}_1\mathbf{Z}_d]^{-1}$and $\mathbf{B}_d = \mathbf{Z}'_d\mathbf{L}_2\mathbf{Z}_d[\mathbf{Z}'_d\mathbf{L}_1\mathbf{Z}_d + \mathbf{Z}'_d\mathbf{L}_2\mathbf{Z}_d]^{-1}$, with $\mathbf{L}_i = \mathbf{I} - \mathbf{P}_{\mathbf{X}_i}$, are the population counterparts of $\hat{\mathbf{B}}_{y_{id}}$ and $\hat{\mathbf{B}}_d$, respectively. The index $i$ in $\mathbf{L}_i$ indicates possibly different auxiliary variables associated with the two samples. Analogous is the expression of $AV(\hat{Y}^{CR}_{2d})$. Furthermore, $AV(\hat{\mathbf{Z}}^{CR}_{id})$ is given by

$$\begin{aligned} AV(\hat{\mathbf{Z}}^{CR}_{id}) &= \mathbf{B}_d AV(\hat{\mathbf{Z}}^R_{1d})\mathbf{B}'_d \\ &\quad + (\mathbf{I} - \mathbf{B}_d)AV(\hat{\mathbf{Z}}^R_{2d})(\mathbf{I} - \mathbf{B}_d)'. \end{aligned} \qquad (10)$$

It is clear from the above that estimates for common and noncommon variables are obtained using the data of only one of the surveys. Further, it is important to note that each sample's calibrated weights incorporate auxiliary information from the other sample. This suggests that this special regression procedure that combines data from the two samples should produce composite estimators (6) and (8) that are more efficient than the regression estimators based on one sample, more so for the common vector variable $\mathbf{z}$ because of the direct correlation of its values from the two samples. This, however, is not necessarily the case. For instance, when the auxiliary variables used in the two surveys are the same, the sample quantities $\mathbf{Z}'_{s_{1d}}\mathbf{L}_{s_1}\mathbf{Z}_{s_{1d}}$ and $\mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_2}\mathbf{Z}_{s_{2d}}$ are estimates of the same population quantity and, therefore, the coefficients $\mathbf{B}_d$ and $\mathbf{I} - \mathbf{B}_d$ are both equal to $(1/2)\mathbf{I}$, so that (10) becomes $AV(\hat{\mathbf{Z}}^{CR}_{1d}) = AV(\hat{\mathbf{Z}}^{CR}_{2d}) = (1/4)[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})]$. It follows then that $AV(\hat{\mathbf{Z}}^{CR}_{1d}) \leq AV(\hat{\mathbf{Z}}^R_{1d})$ only if $AV(\hat{\mathbf{Z}}^R_{2d}) \leq AV(\hat{\mathbf{Z}}^R_{1d})$. In the case of simple random sampling without replacement (SRSWOR) for both surveys with sampling fractions $f_i = n_i/N$ and with the ratio of the finite population corrections $1 - f_1$ and $1 - f_2$ approximately equal to 1, it can be shown that $AV(\hat{\mathbf{Z}}^{CR}_{1d}) \leq AV(\hat{\mathbf{Z}}^R_{1d})$ only if $n_2 \geq n_1/3$, and $AV(\hat{Y}^{CR}_{1d}) \leq AV(\hat{Y}^R_{1d})$ only if $n_2 \geq n_1$. When $n_1 = n_2$, $AV(\hat{Y}^{CR}_{1d}) = AV(\hat{Y}^R_{1d})$, while $AV(\hat{\mathbf{Z}}^{CR}_{1d}) = (1/2)AV(\hat{\mathbf{Z}}^R_{1d})$.

As this particular situation attests, the multivariate sample coefficient $\hat{\mathbf{B}}_d$ generated implicitly by the regression procedure does not account for any difference in sample size between the two samples, though it incorporates the relative effect of regression fit in $\hat{\mathbf{Z}}^R_{1d}$ and $\hat{\mathbf{Z}}^R_{2d}$. The same is true, but not as apparent, when the auxiliary vectors used in the two surveys are not identical.

To account for the differential in effective sample size between two samples having arbitrary sampling designs, we can adapt to the present context a modification of the GREG procedure suggested in Merkouris (2004) for combining information at the population level. It involves the replacement of the quadratic forms $\mathbf{Z}'_{s_{id}}\mathbf{L}_{s_i}\mathbf{Z}_{s_{id}}$ by the respective mean forms $(1/\tilde{n}_i)\mathbf{Z}'_{s_{id}}\mathbf{L}_{s_i}\mathbf{Z}_{s_{id}}$, where $\tilde{n}_i = n_i/d_i$ are the effective sample sizes — $d_i$ denoting design effects. To this end, all is needed is the scaling adjustment of the entries of the weighting matrix $\mathbf{\Lambda}_{s_i}$ by $1/\tilde{n}_i$. The composite regression estimators (6) and (8) are affected by this adjustment only through the regression coefficient $\hat{\mathbf{B}}_d$, which can now be written as $\hat{\mathbf{B}}_d = \phi\mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_2}\mathbf{Z}_{s_{2d}}[(1 - \phi)\mathbf{Z}'_{s_{1d}}\mathbf{L}_{s_1}\mathbf{Z}_{s_{1d}} + \phi\mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_2}\mathbf{Z}_{s_{2d}}]^{-1}$, where $\phi = \tilde{n}_1/(\tilde{n}_1 + \tilde{n}_2)$. In fact, an equivalent adjustment can be made through scaling $\mathbf{\Lambda}_{s_1}$ by $(1-\phi)$ and $\mathbf{\Lambda}_{s_2}$ by $\phi$. A least-squares characterization of $\hat{\mathbf{B}}_d$ and related efficiency considerations are as in Merkouris (2004).

It is interesting to see if the adjusted composite estimators (6) and (8) are more efficient than their single-sample components. An exact analytical result is furnished by the following proposition; the proof is given in the Appendix.

**Proposition 1** *(a) Suppose that $\mathbf{1} = \mathbf{X}_i\mathbf{h}_i$, for constant $p_i$-vectors $\mathbf{h}_i$. Assume simple random sampling without replacement with sampling fractions $f_i = n_i/N$ such that $(1 - f_1)/(1 - f_2) \approx 1$. Then the following inequalities hold.*

$$AV(\hat{\mathbf{Z}}^{CR}_{id}) < AV(\hat{\mathbf{Z}}^R_{id}), \qquad AV(\hat{Y}^{CR}_{id}) < AV(\hat{Y}^R_{id}). \quad (11)$$

*Furthermore, when $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the same variables,*

$$\begin{aligned} AV(\hat{\mathbf{Z}}^R_{id})[AV(\hat{\mathbf{Z}}^{CR}_{id})]^{-1} &= [(n_1 + n_2)/n_i]\mathbf{I}, \\ AV(\hat{Y}^R_{id})[AV(\hat{Y}^{CR}_{id})]^{-1} &< (n_1 + n_2)/n_i. \end{aligned} \qquad (12)$$

*(b) The results in (a) hold also under Bernoulli sampling with probabilities of inclusion $\pi_i = f_i (= n_i/N)$ satisfying $(1 - f_1)/(1 - f_2) \approx 1$.*

The inequality $AV(\hat{\mathbf{Z}}_{id}^{CR}) < AV(\hat{\mathbf{Z}}_{id}^{R})$ holds in the partial ordering of nonnegative definite matrices, and therefore it also holds for any linear combination of the components of each of the estimators involved. The condition $\mathbf{1} = \mathbf{X}\mathbf{h}$ of a projection estimator in the above theorem is customarily satisfied in surveys that use GREG estimation. The exact result in (12), when the same auxiliary vector is used in both surveys, shows that the efficiency of $\hat{\mathbf{Z}}_{id}^{CR}$ relative to $\hat{\mathbf{Z}}_{id}^{R}$ (component-wise) can be substantial, e.g., 100% if $n_1 = n_2$. For $\hat{Y}_{id}^{CR}$, which borrows strength indirectly through the correlation of $y_i$ with $\mathbf{z}$, the gain is smaller. With the population quantity $\mathbf{B}_d$, too, incorporating the adjustments $(1-\phi)$ and $\phi$, and under the conditions of Proposition 1, it can be shown that $\mathbf{B}_{y_{id}}(\mathbf{I}-\mathbf{B}_d) = AC(\hat{Y}_{id}^{R}, \hat{\mathbf{Z}}_{1d}^{R})[AV(\hat{\mathbf{Z}}_{1d}^{R}) + AV(\hat{\mathbf{Z}}_{2d}^{R})]^{-1}$ and $\mathbf{B}_d = AV(\hat{\mathbf{Z}}_{2d}^{R})[AV(\hat{\mathbf{Z}}_{1d}^{R}) + AV(\hat{\mathbf{Z}}_{2d}^{R})]^{-1}$, so that these are the optimal (variance minimizing) coefficients in (9) and ( 10). Under more general settings, the efficiency gain will be somewhat smaller, as the coefficients $\mathbf{B}_{y_{id}}(\mathbf{I}-\mathbf{B}_d)$ and $\mathbf{B}_d$ will only be approximations of the optimal ones. Also, it is clear from (9) that the efficiency of $\hat{Y}_{id}^{CR}$ depends on the strength of correlation between $y_i$ and $\mathbf{z}$.

If we choose to combine information on a subset of the common variables, then for the rest we derive two domain estimators, as in (6), and it would be then beneficial to combine them in some way. A sensible combination involves weighting the individual composite estimators proportionally to the effective size of the associated sample. Such combination would give the composite estimator $\hat{Y}_{d}^{CR} = \phi\hat{Y}_{1d}^{CR} + (1-\phi)\hat{Y}_{2d}^{CR}$, where $\phi = \tilde{n}_1/(\tilde{n}_1 + \tilde{n}_2)$. Under the conditions of Proposition 1, the approximate variance of $\hat{Y}_{d}^{CR}$ can be derived without difficulty as $AV(\hat{Y}_{d}^{CR}) = \phi^2 AV(\hat{Y}_{1d}^{R}) + (1-\phi)^2 AV(\hat{Y}_{2d}^{R}) - \mathbf{a}[AV(\hat{\mathbf{Z}}_{1d}^{R}) + (AV(\hat{\mathbf{Z}}_{2d}^{R})]^{-1}\mathbf{a}'$, where $\mathbf{a} = \phi AC(\hat{Y}_{1d}^{R}, \hat{\mathbf{Z}}_{1d}^{R}) - (1-\phi)AC(\hat{Y}_{2d}^{R}, \hat{\mathbf{Z}}_{2d}^{R})$. Clearly, unless $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the same variables (implying $\mathbf{a} = \mathbf{0}$), the variance of $\hat{Y}_{d}^{CR}$ is strictly smaller than the variance of the composite $\phi\hat{Y}_{1d}^{R} + (1-\phi)\hat{Y}_{2d}^{R}$ of the initial independent GREG estimators — but not by much, in view of $\mathbf{a}$. The computation of the composite $\hat{Y}_{d}^{CR}$ can be incorporated into the GREG composite estimation procedure without difficulty, whereas an optimal linear combination of the dependent estimators $\hat{Y}_{1d}^{CR}$ and $\hat{Y}_{2d}^{CR}$ would not be practical, and probably not considerably more efficient.

The results of this section can be generalized to any number of domains. For example, for the complementary domains $U_d$ and $U_{\bar{d}}$ the matrices $\mathbf{Z}_{s_{id}}$ in the setup (5) will be augmented to $(\mathbf{Z}_{s_{id}}, \mathbf{Z}_{s_{i\bar{d}}})$. Nothing changes formally in the expressions above if the index $d$ is to simply indicate that for various domains the information on $\mathbf{z}$ from the two samples is combined at the domain level, and that ( 6) and (8) give estimates for $\mathbf{t}_{y_{id}}$ and $\mathbf{t}_{\mathbf{z}_d}$ for each of these domains.

### 3.2 Regressing and combining information at the $U_d$ level

We now introduce the variant of the regression set up (5)

$$\boldsymbol{\mathcal{X}}_{s_d} = \begin{pmatrix} \mathbf{X}_{s_{1d}} & \mathbf{0} & \mathbf{Z}_{s_{1d}} \\ \mathbf{0} & \mathbf{X}_{s_{2d}} & -\mathbf{Z}_{s_{2d}} \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \mathbf{t}_{\mathbf{x}_{1d}} \\ \mathbf{t}_{\mathbf{x}_{2d}} \\ \mathbf{0} \end{pmatrix}, \quad (13)$$

whereby regression on $\mathbf{x}_1$ and $\mathbf{x}_2$ is carried out at the domain level. This yields the composite domain estimators

$$\begin{aligned} \check{Y}_{1d}^{CR} &= \check{Y}_{1d}^{R} + \check{\mathbf{B}}_{y_{1d}}(\mathbf{I} - \check{\mathbf{B}}_d)[\check{\mathbf{Z}}_{2d}^{R} - \check{\mathbf{Z}}_{1d}^{R}], \\ \check{Y}_{2d}^{CR} &= \check{Y}_{2d}^{R} - \check{\mathbf{B}}_{y_{2d}}\check{\mathbf{B}}_d[\check{\mathbf{Z}}_{2d}^{R} - \check{\mathbf{Z}}_{1d}^{R}], \end{aligned} \quad (14)$$

and

$$\check{\mathbf{Z}}_{1d}^{CR} = \check{\mathbf{Z}}_{2d}^{CR} = \check{\mathbf{B}}_d\check{\mathbf{Z}}_{1d}^{R} + (\mathbf{I} - \check{\mathbf{B}}_d)\check{\mathbf{Z}}_{2d}^{R}, \quad (15)$$

where $\check{Y}_{1d}^{R} = \hat{Y}_{id} + \mathbf{Y}'_{s_{id}}\boldsymbol{\Lambda}_{s_{id}}\mathbf{X}_{s_{id}}(\mathbf{X}'_{s_{id}}\boldsymbol{\Lambda}_{s_{id}}\mathbf{X}_{s_{id}})^{-1}[\mathbf{t}_{\mathbf{x}_{id}} - \hat{\mathbf{X}}_{id}]$ and $\check{\mathbf{Z}}_{id}^{R} = \hat{\mathbf{Z}}_{id} + \mathbf{Z}'_{s_{id}}\boldsymbol{\Lambda}_{s_{id}}\mathbf{X}_{s_{id}}(\mathbf{X}'_{s_{id}}\boldsymbol{\Lambda}_{s_{id}}\mathbf{X}_{s_{id}})^{-1}[\mathbf{t}_{\mathbf{x}_{id}} - \hat{\mathbf{X}}_{id}]$ are GREG domain estimators using auxiliary data only from $U_d$, and $\check{\mathbf{B}}_{y_{1d}} = \mathbf{Y}'_{s_{id}}\mathbf{L}_{s_{id}}\mathbf{Z}_{s_{id}}[\mathbf{Z}'_{s_{id}}\mathbf{L}_{s_{id}}\mathbf{Z}_{s_{id}}]^{-1}$, $\check{\mathbf{B}}_d = \mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_{2d}}\mathbf{Z}_{s_{2d}}[\mathbf{Z}'_{s_{1d}}\mathbf{L}_{s_{1d}}\mathbf{Z}_{s_{1d}} + \mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_{2d}}\mathbf{Z}_{s_{2d}}]^{-1}$ with $\mathbf{L}_{s_{id}} = (1/\tilde{n}_i)\boldsymbol{\Lambda}_{s_i}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{s_{id}}})$.

The approximate design variance of $\check{Y}_{1d}^{CR}$ is given by

$$\begin{aligned} AV(\check{Y}_{1d}^{CR}) = {} & AV(\check{Y}_{1d}^{R}) \\ & +\mathbf{B}_{y_{1d}}(\mathbf{I} - \mathbf{B}_d)[AV(\check{\mathbf{Z}}_{1d}^{R}) + AV(\check{\mathbf{Z}}_{2d}^{R})] \\ & \times(\mathbf{I} - \mathbf{B}_d)'\mathbf{B}'_{y_{1d}} - 2\mathbf{B}_{y_{1d}}(\mathbf{I} - \mathbf{B}_d) \\ & \times[AC(\check{Y}_{1d}^{R}, \check{\mathbf{Z}}_{1d}^{R})]', \end{aligned} \quad (16)$$

where $\mathbf{B}_{y_{1d}}$ and $\mathbf{B}_d$ are the population counterparts of $\check{\mathbf{B}}_{y_{1d}}$ and $\check{\mathbf{B}}_d$, respectively. $AV(\check{\mathbf{Z}}_{1d}^{CR})$ and $AV(\check{\mathbf{Z}}_{2d}^{CR})$ are given by

$$\begin{aligned} AV(\check{\mathbf{Z}}_{id}^{CR}) = {} & \mathbf{B}_d AV(\check{\mathbf{Z}}_{1d}^{R})\mathbf{B}'_d \\ & +(\mathbf{I} - \mathbf{B}_d)AV(\check{\mathbf{Z}}_{2d}^{R})(\mathbf{I} - \mathbf{B}_d)'. \end{aligned} \quad (17)$$

Results identical to those of Proposition 1 hold for the estimators (14) and (15); the proof is similar to that of Proposition 1. The estimators (14) and (15) are expected to be highly efficient because the regression on $\mathbf{x}_1$ and $\mathbf{x}_2$ is at the domain level. The above results can be generalized to any number of domains. Note that population-level estimators can be obtained from the domain estimators additively.

### 3.3 Incorporating $N_d$ in the regression and combining information at the $U_d$ level

The domain totals $\mathbf{t}_{\mathbf{x}_{id}}$ used in (13) may not be readily available or may be of questionable quality, especially for very small domains. Moreover, the domain sample sizes may be very small or the number of auxiliary variables may be too large for the available domain sample sizes; this can cause significant bias and inflation of the variance of the derived composite estimators. It may then be more sensible to use as auxiliary information at the domain level only the domain size, using the setup

$$\boldsymbol{\mathcal{X}}_s = \begin{pmatrix} \boldsymbol{\chi}_{s_1} & \mathbf{0} & \mathbf{Z}_{s_{1d}} \\ \mathbf{0} & \boldsymbol{\chi}_{s_2} & -\mathbf{Z}_{s_{2d}} \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \mathbf{t}_{\boldsymbol{\chi}_1} \\ \mathbf{t}_{\boldsymbol{\chi}_2} \\ \mathbf{0} \end{pmatrix}, \tag{18}$$

where $\boldsymbol{\chi}_{s_i} = (\mathbf{X}_{s_i}\ \mathbf{1}_{s_{id}})$ and $\mathbf{t}_{\boldsymbol{\chi}_i} = (\mathbf{t}'_{\mathbf{x}_i}, N_d)'$. This yields the composite domain estimators

$$\begin{aligned} \breve{Y}^{CR}_{1d} &= \breve{Y}^{R}_{1d} + \breve{\mathbf{B}}_{y_{1d}}(\mathbf{I} - \breve{\mathbf{B}}_d)[\breve{\mathbf{Z}}^R_{2d} - \breve{\mathbf{Z}}^R_{1d}], \\ \breve{Y}^{CR}_{2d} &= \breve{Y}^{R}_{2d} - \breve{\mathbf{B}}_{y_{2d}}\breve{\mathbf{B}}_d[\breve{\mathbf{Z}}^R_{2d} - \breve{\mathbf{Z}}^R_{1d}], \end{aligned} \tag{19}$$

and

$$\breve{\mathbf{Z}}^{CR}_{1d} = \breve{\mathbf{Z}}^{CR}_{2d} = \breve{\mathbf{B}}_d\breve{\mathbf{Z}}^R_{1d} + (\mathbf{I} - \breve{\mathbf{B}}_d)\breve{\mathbf{Z}}^R_{2d}, \tag{20}$$

where $\breve{Y}^R_{id}$, $\breve{\mathbf{Z}}^R_{id}$ are GREG domain estimators based on $\boldsymbol{\chi}_{s_i} = (\mathbf{X}_{s_i}\ \ \mathbf{1}_{s_{id}})$ and $\mathbf{t}_{\boldsymbol{\chi}_i} = (\mathbf{t}'_{\mathbf{x}_i}, N_d)'$, and $\breve{\mathbf{B}}_{y_{id}} = \mathbf{Y}'_{s_{id}}\boldsymbol{\mathcal{L}}_{s_i}\mathbf{Z}_{s_{id}}[\mathbf{Z}'_{s_{id}}\boldsymbol{\mathcal{L}}_{s_i}\mathbf{Z}_{s_{id}}]^{-1}$, $\breve{\mathbf{B}}_d = \mathbf{Z}'_{s_{2d}}\boldsymbol{\mathcal{L}}_{s_2}\mathbf{Z}_{s_{2d}}[\mathbf{Z}'_{s_{1d}}\boldsymbol{\mathcal{L}}_{s_1}\mathbf{Z}_{s_{1d}} + \mathbf{Z}'_{s_{2d}}\boldsymbol{\mathcal{L}}_{s_2}\mathbf{Z}_{s_{2d}}]^{-1}$ with $\boldsymbol{\mathcal{L}}_{s_i} = (1/\tilde{n}_i)\boldsymbol{\Lambda}_{s_i}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\chi}_{s_i}})$. We ensure, of course, that $(\boldsymbol{\chi}_{s_i}\ \mathbf{Z}_{s_{id}})$ is of full rank $p_i + 1 + q$.

The approximate design variance of $\breve{Y}^{CR}_{1d}$ is given by

$$\begin{aligned} AV(\breve{Y}^{CR}_{1d}) \;=\; & AV(\breve{Y}^{R}_{1d}) \\ & + \mathbf{B}_{y_{1d}}(\mathbf{I} - \mathbf{B}_d)[AV(\breve{\mathbf{Z}}^R_{1d}) + AV(\breve{\mathbf{Z}}^R_{2d})] \\ & \times(\mathbf{I} - \mathbf{B}_d)'\mathbf{B}'_{y_{1d}} - 2\mathbf{B}_{y_{1d}}(\mathbf{I} - \mathbf{B}_d) \\ & \times[AC(\breve{Y}^R_{1d}, \breve{\mathbf{Z}}^R_{1d})]', \end{aligned} \tag{21}$$

where $\mathbf{B}_{y_{1d}}$ and $\mathbf{B}_d$ are the population counterparts of $\breve{\mathbf{B}}_{y_{1d}}$ and $\breve{\mathbf{B}}_d$, respectively. $AV(\breve{\mathbf{Z}}^{CR}_{1d})$ and $AV(\breve{\mathbf{Z}}^{CR}_{2d})$ are given by

$$\begin{aligned} AV(\breve{\mathbf{Z}}^{CR}_{id}) \;=\; & \mathbf{B}_d AV(\breve{\mathbf{Z}}^R_{1d})\mathbf{B}'_d \\ & + (\mathbf{I} - \mathbf{B}_d)AV(\breve{\mathbf{Z}}^R_{2d})(\mathbf{I} - \mathbf{B}_d)'. \end{aligned} \tag{22}$$

Results identical to those of Proposition 1 hold for the estimators (19) and (20); the proof is similar to that of Proposition 1. Here again, a generalization to more than one domain is straightforward.

The three composite estimators (8), (15) (20) are compared with respect to their approximate variance under the conditions of the following theorem; the proof is given in the Appendix.

**Theorem 1** *(a) Suppose that $\mathbf{1} = \mathbf{X}_i\mathbf{h}_i$, for constant $p_i$-vectors $\mathbf{h}_i$. Assume simple random sampling without replacement with sampling fractions $f_i = n_i/N$ such that $(1 - f_1)/(1 - f_2) \approx 1$. Then the following inequalities hold.*

$$AV(\check{\mathbf{Z}}^{CR}_{id}) \leq AV(\breve{\mathbf{Z}}^{CR}_{id}) \leq AV(\hat{\mathbf{Z}}^{CR}_{id}). \tag{23}$$

*(b) Under Bernoulli sampling with $\pi_i = f_i$ as in (a), the inequality $AV(\check{\mathbf{Z}}^{CR}_{id}) \leq AV(\breve{\mathbf{Z}}^{CR}_{id})$ holds under the condition $\mathbf{1} = \mathbf{X}_i\mathbf{h}_i$; the other parts of (23) hold without this condition.*

The inequality (23) shows that under the conditions of Theorem 1 more domain-specific information about the auxiliary variables $\mathbf{x}_1$ and $\mathbf{x}_2$ leads to more efficient composite estimators of the domain total $\mathbf{t}_{\mathbf{z}_d}$. Similar results are expected to hold under general designs.

## 4 Summary and discussion

Analytical results for simple random sampling and Bernoulli sampling show that extending the generalized regression procedure so as to combine comparable information from two surveys at the domain level improves the precision of domain estimates, substantially for common survey variables but less so for noncommon variables. The precision gain increases with the use of more domain-specific auxiliary information.

For an empirical study of the proposed estimation method in a complex survey context, a two-sample situation was created by splitting the sample of the Canadian Labour Force Survey (LFS) by rotation group into two subsamples, $s_1$ and $s_2$, of three rotations each. The six rotations that comprise the LFS are independent samples (of approximately the same size) of members of private households drawn with a stratified multistage design from an area frame. For the purposes of the study, the two main categories of the Labour Force status, i.e., "employed" and "unemployed", were chosen as "common" variables to the two subsamples. Regression was carried out employing the same calibration scheme for $s_1$ and $s_2$ involving seven age groups by sex, and incorporating the two common variables simultaneously. Four small geographic areas in each of the two provinces were used as study domains. The one-sample domain estimators $\hat{Z}^R_d$ and $\breve{Z}^R_d$ and their composite counterparts $\hat{Z}^{CR}_d$ and $\breve{Z}^{CR}_d$ (for scalar characteristics) were compared with respect to their estimated (by the Jackknife method) variances. This study provided a quantification of the resulting efficiency gains for estimated totals of common and noncommon binary variables. In particular, in a comparison of a regression procedure that uses only auxiliary information at the domain level with a regression procedure that only combines data from two surveys, the study has shown that the former is more efficient for the most prevalent of two common characteristics but the latter is more efficient for the less prevalent one and for associated rates. The total effect of using auxiliary information and combining information on common variables through regression is an impressively efficient domain estimation, more so for the common variables. These empirical results were not based on repeated samples, and may therefore be regarded as strongly suggestive but not conclusive. Details of this study can be found in Merkouris (2006).

It has been assumed that domains that are of interest to one of the surveys are identifiable in the data file of the other survey. When this is not the case, combining information on common variables is still possible but at the next higher subpopulation level identifiable in both files, which may be the entire population. For general designs, the efficiency of the resulting composite domain estimators relative to the one-sample domain GREG estimators will depend on that level. A situation where it may be preferable to combine information on some of the common variables at intermediate levels identifiable in both samples (e.g., strata containing the domains of interest) involves very small domains in which those variables represent rare characteristics. It is to be noted that combining information at levels other than the domains of interest does not lead to composite domain estimators that have the form of a weighted average.

It was shown in Section 3 that the proposed GREG procedure may conveniently handle more than one common variable and more than one domain at once; the various domains need not be mutually exclusive and exhaustive. This procedure generates a single set of calibrated weights for each survey that can be used to produce a composite estimate for any variable of interest, common or noncommon, and at any level, thereby preserving each survey's internal consistency of estimates. With such a unified approach to estimation of any parameter of interest, it is sensible to combine information also at the population level — by augmenting the regression matrix (5) or (13) or (18) with the submatrix $(\mathbf{Z}_{s_1}, -\mathbf{Z}_{s_2})'$. This will be redundant if the domains included in the procedure are exhaustive.

The number of common variables for which we seek domain estimates may be so large so as to make the regression procedure too cumbersome or lead to unstable estimates. A large number of domains of interest may have the same effects. In such situations, it may be more appropriate to carry out separate GREG procedures with a subset of domains and (or) a subset of the common variables. With such an approach we forgo a unified estimation system, as is rather customary in domain estimation based on a single sample, but we obtain more stable estimates and have more flexibility in the use of the auxiliary variables $\mathbf{x_1}$ and $\mathbf{x_2}$. In particular, noting that population-level controls are ineffective for domain estimation, it is prudent to use only controls that are available at the level of subpopulation (domain or higher) at which the information from $s_1$ and $s_2$ is combined.

In the special situation when all variables are common between the two surveys (as when one of the surveys is supplementary of the other) and consistency of estimates is required, we may choose to combine information on a subset of key common variables. For the rest we derive two domain estimators, as in (6), (14) or (19), which can be combined as already described in Section 3.1 for the estimators in (6).

As already noted, the use of domain-level control totals in the GREG procedure may be limited to domain sizes because domain totals for some or all the components of an auxiliary vector $\mathbf{x}$ may be unavailable, or because the domain sample size is too small. However, even domain sizes may not be available for some small domains, most likely not for non-geographic domains. On the other hand, combining information on common variables in such domains is always possible. This is a great advantage of this approach to domain estimation, all the more so considering that in some situations small domains may not be adequately amenable to traditional model-based techniques. Moreover, calibrating to the size of small domains may result in loss of efficiency for small proportions within these domains due to small sample count, and, for the same reason, it may introduce some bias to domain estimators. In contrast, combining information on common variables at the domain level essentially increases the effective domain sample size.

If no auxiliary variables are used, the regression matrix (5) reduces to $(\mathbf{Z}_{s_{1d}}, -\mathbf{Z}_{s_{2d}})'$ and the composite GREG estimators (6) and (8) are given in terms of the HT estimators from the two samples. For example, the explicit form of $\hat{Y}^{CR}_{1d}$ will be $\hat{Y}^{CR}_{1d} = \hat{Y}_{1d} + (1-\phi)\mathbf{Y}'_{s_{1d}}\mathbf{\Lambda}_{s_1}\mathbf{Z}_{s_{1d}}[(1-\phi)\mathbf{Z}'_{s_{1d}}\mathbf{\Lambda}_{s_1}\mathbf{Z}_{s_{1d}} + \phi\mathbf{Z}'_{s_{2d}}\mathbf{\Lambda}_{s_2}\mathbf{Z}_{s_{2d}}]^{-1}[\hat{\mathbf{Z}}_{2d} - \hat{\mathbf{Z}}_{1d}]$. In this form, the beneficial effect of the increased effective sample size in enhancing the stability of the regression coefficient is evident.

Extensions of the analytic results of Section 3 to general sampling designs appear to be intractable, except for direct extentions to stratified sampling with either separate or combined regression for domains that cut across strata, as done in Merkouris (2004) for population-level combination.

Finally, although situations involving more than two surveys with overlapping content from the same population are rather unusual, a suitable generalization of the proposed procedure (following Merkouris (2004)) is easy.

## 5 Appendix: Proofs

*Proof of Proposition 1.* The proof will be given for $\hat{\mathbf{Z}}^R_{1d}$ and $\hat{Y}^R_{1d}$; the proof for $\hat{\mathbf{Z}}^R_{2d}$ and $\hat{Y}^R_{2d}$ is similar. (a) First, the vector of population residuals corresponding to $\hat{\mathbf{Z}}^R_{id}$ is $\mathbf{E}_{id} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{Z}_d$, where $\mathbf{P}_{\mathbf{X}_i} = \mathbf{X}_i(\mathbf{X}'_i\mathbf{X}_i)^{-1}\mathbf{X}'_i$. Using a matrix formulation of a standard result (see, e.g., Särndal et al. 1992, p. 235), the approximate design variance of $\hat{\mathbf{Z}}^R_{id}$ is given by

$$AV(\hat{\mathbf{Z}}^R_{id}) = \mathbf{E}'_{id}\mathbf{\Lambda}^\circ_i\mathbf{E}_{id} = \mathbf{Z}'_d(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{\Lambda}^\circ(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{Z}_d,$$

where $\mathbf{\Lambda}^\circ_i$ is a nonnegative definite matrix whose $kl$-th entry is (dropping the index $i$ for notational simplicity) $(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l$, $(\pi_{kk} \equiv \pi_k)$. For simple random sampling with sampling fraction $f_i = n_i/N$, it can be easily shown that $\mathbf{\Lambda}^\circ_i = \lambda^\circ_i(\mathbf{I} - \mathbf{P_1})$, where $\lambda^\circ_i = N^2(1-f_i)/[n_i(N-1)]$ and $\mathbf{P_1} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$. Then, it follows from $\mathbf{1} = \mathbf{X}_i\mathbf{h}_i$ that $\mathbf{P}_{\mathbf{X}_i}\mathbf{P_1} = \mathbf{P_1}$ and thus $AV(\hat{\mathbf{Z}}^R_{id}) = \lambda^\circ_i\mathbf{Z}'_d(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{Z}_d = \lambda^\circ_i\mathbf{Z}'_d\mathbf{L}_i\mathbf{Z}_d$. Since the columns of $\mathbf{Z}_d$ are independent of the columns of $\mathbf{X}_i$, the matrices $\mathbf{Z}'_d\mathbf{L}_i\mathbf{Z}_d$ are nonsingular (see Seber 1997, p.65). It follows then that $\mathbf{B}_d = AV(\hat{\mathbf{Z}}^R_{2d})[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})]^{-1}$ under the condition $(1-f_1)/(1-f_2) \approx 1$. Now, expression (10) can be rewritten as $AV(\hat{\mathbf{Z}}^{CR}_{1d}) = AV(\hat{\mathbf{Z}}^R_{1d})[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})]^{-1}AV(\hat{\mathbf{Z}}^R_{2d}) = AV(\hat{\mathbf{Z}}^R_{1d}) - AV(\hat{\mathbf{Z}}^R_{1d})[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})]^{-1}AV(\hat{\mathbf{Z}}^R_{1d})$, which shows that $AV(\hat{\mathbf{Z}}^{CR}_{1d}) < AV(\hat{\mathbf{Z}}^R_{1d})$.

Similarly, (9) can take the form $AV(\hat{Y}^{CR}_{1d}) = AV(\hat{Y}^R_{1d}) - AC(\hat{Y}^R_{1d}, \hat{\mathbf{Z}}^R_{1d})[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})]^{-1} \times (AC(\hat{Y}^R_{1d}, \hat{\mathbf{Z}}^R_{1d}))'$, which shows that $AV(\hat{Y}^{CR}_{1d}) < AV(\hat{Y}^R_{1d})$.

Now, $\mathbf{X}_1 = \mathbf{X}_2$ implies $\mathbf{I} - \mathbf{P}_{\mathbf{X}_1} = \mathbf{I} - \mathbf{P}_{\mathbf{X}_2}$ and $\mathbf{B}_d = \phi\mathbf{I}$, $\phi = n_1/(n_1+n_2)$. Therefore, $AV(\hat{\mathbf{Z}}^{CR}_{1d}) = [\lambda^\circ_1\phi^2 + \lambda^\circ_2(1-\phi)^2]\mathbf{Z}'_d(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Z}'_d$. Also, $AV(\hat{\mathbf{Z}}^R_{1d}) = \lambda^\circ_1\mathbf{Z}'_d(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Z}'_d$, and thus $AV(\hat{\mathbf{Z}}^R_{1d})[AV(\hat{\mathbf{Z}}^{CR}_{1d})]^{-1} = [\lambda^\circ_1/[\lambda^\circ_1\phi^2 + \lambda^\circ_2(1-\phi)^2]]\mathbf{I} = [(n_1+n_2)/n_1]\mathbf{I}$ under the condition $(1-f_1)/(1-f_2) \approx 1$.

Also, $\mathbf{B}_{y_{1d}}(\mathbf{I} - \mathbf{B}_d) = (1-\phi)AC(\hat{Y}^R_{1d}, \hat{\mathbf{Z}}^R_{1d})[AV(\hat{\mathbf{Z}}^R_{1d})]^{-1}$ and thus $AV(\hat{Y}^{CR}_{1d}) = AV(\hat{Y}^R_{1d}) + [\phi^2 - 1 + \lambda^\circ_2(1-\phi)^2/\lambda^\circ_1]AC(\hat{Y}^R_{1d}, \hat{\mathbf{Z}}^R_{1d})[AV(\hat{\mathbf{Z}}^R_{1d})]^{-1}(AC(\hat{Y}^R_{1d}, \hat{\mathbf{Z}}^R_{1d}))'$. By the Cauchy-Schwarz inequality $AV(\hat{Y}^{CR}_{1d}) > AV(\hat{Y}^R_{1d}) + [\phi^2 - 1 + \lambda^\circ_2(1-\phi)^2/\lambda^\circ_1]AV(\hat{Y}^R_{1d}) = [[\lambda^\circ_1\phi^2 + \lambda^\circ_2(1-\phi)^2]/\lambda^\circ_1]AV(\hat{Y}^R_{1d})$. It follows then that $AV(\hat{Y}^R_{1d})[AV(\hat{Y}^{CR}_{1d})]^{-1} < (n_1+n_2)/n_1$.

(b) The proof is as in (a).

*Proof of Theorem 1.* It suffices to give the proof for $\hat{\mathbf{Z}}^{CR}_{1d}$, $\breve{\mathbf{Z}}^{CR}_{1d}$ and $\check{\mathbf{Z}}^{CR}_{1d}$. (a) It was shown in the proof of Proposition 1 that $AV(\hat{\mathbf{Z}}^{CR}_{1d}) = AV(\hat{\mathbf{Z}}^R_{1d})[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})]^{-1}AV(\hat{\mathbf{Z}}^R_{2d})$. Analogous expressions can be derived for $AV(\breve{\mathbf{Z}}^{CR}_{1d})$ and $AV(\check{\mathbf{Z}}^{CR}_{1d})$. Now, noticing that $AV(\hat{\mathbf{Z}}^R_{1d})[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})]^{-1}AV(\hat{\mathbf{Z}}^R_{2d}) = [(AV(\hat{\mathbf{Z}}^R_{2d}))^{-1}[AV(\hat{\mathbf{Z}}^R_{1d}) + AV(\hat{\mathbf{Z}}^R_{2d})](AV(\hat{\mathbf{Z}}^R_{1d}))^{-1}]^{-1} = [(AV(\hat{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\hat{\mathbf{Z}}^R_{2d}))^{-1}]^{-1}$, we can write

$$AV(\hat{\mathbf{Z}}^{CR}_{1d}) = [(AV(\hat{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\hat{\mathbf{Z}}^R_{2d}))^{-1}]^{-1}.$$

Similarly, $AV(\breve{\mathbf{Z}}^{CR}_{1d}) = [(AV(\breve{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\breve{\mathbf{Z}}^R_{2d}))^{-1}]^{-1}$ and $AV(\check{\mathbf{Z}}^{CR}_{1d}) = [(AV(\check{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\check{\mathbf{Z}}^R_{2d}))^{-1}]^{-1}$. For $\hat{\mathbf{Z}}^{CR}_{1d}$ and $\breve{\mathbf{Z}}^{CR}_{1d}$ we obtain

$$\begin{aligned} AV(\hat{\mathbf{Z}}^{CR}_{1d}) - AV(\breve{\mathbf{Z}}^{CR}_{1d}) &= [(AV(\hat{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\hat{\mathbf{Z}}^R_{2d}))^{-1}]^{-1} \\ &\quad -[(AV(\breve{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\breve{\mathbf{Z}}^R_{2d}))^{-1}]^{-1}. \end{aligned}$$

Since the matrices $AV(\hat{\mathbf{Z}}^{CR}_{1d})$ and $AV(\breve{\mathbf{Z}}^{CR}_{1d})$ are nonnegative definite and nonsingular, they are positive definite. It can be shown (see Merkouris 2006) that $AV(\breve{\mathbf{Z}}^R_{id}) \leq AV(\hat{\mathbf{Z}}^R_{id})$, which by a suitable result on inverses of such matrices (Harville 1997, p. 434) implies $(AV(\breve{\mathbf{Z}}^R_{id}))^{-1} \geq (AV(\hat{\mathbf{Z}}^R_{id}))^{-1}$, so that

$$(AV(\breve{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\breve{\mathbf{Z}}^R_{2d}))^{-1} \geq (AV(\hat{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\hat{\mathbf{Z}}^R_{2d}))^{-1}.$$

By applying again the abovementioned result we obtain

$$\begin{aligned} &[(AV(\hat{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\hat{\mathbf{Z}}^R_{2d}))^{-1}]^{-1} \\ \geq\ &[(AV(\breve{\mathbf{Z}}^R_{1d}))^{-1} + (AV(\breve{\mathbf{Z}}^R_{2d}))^{-1}]^{-1} \end{aligned}$$

and, hence,

$$AV(\hat{\mathbf{Z}}^{CR}_{1d}) \geq AV(\breve{\mathbf{Z}}^{CR}_{1d}).$$

The proof of $AV(\breve{\mathbf{Z}}^{CR}_{1d}) \geq AV(\check{\mathbf{Z}}^{CR}_{1d})$ is as above.

(b) The proof is as in (a).

## References

Harville, D.A. (1997), *Matrix Algebra From A Statistician's Perspective*. Springer-Verlang, New York.

Marker, D.A. (2001), "Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators," *Survey Methodology Journal*, **27**, 183–188.

Merkouris, T. (2004), "Combining independent regression estimators from multiple surveys," *Journal of the American Statistical Association*, **99**, 1131–1139.

Merkouris, T. (2006), "Efficient small-domain estimation by combining information from multiple surveys through regression," Working paper, HSMD-2006-007E, Statistics Canada.

Rao, J.N.K. (2003), *Small Area Estimation*, Wiley.

Renssen, R.H., and Nieuwenbroek, N.J. (1997), "Aligning estimates for common variables in two or more sample surveys," *Journal of the American Statistical Association*, **92**, 368–375.

Särndal, C.E., Swensson, B., and Wretman, J.H. (1992), *Model-Assisted Survey Sampling*, New York: Springer-Verlang.

Seber, G.A.F. (1977), *Linear Regression Analysis*, Wiley.

Zieschang, K.D. (1990), "Sample weighting methods and estimation of totals in the consumer expenditures survey," *Journal of the American Statistical Association*, **85**, 986–1001.

Wu, C. (2004), "Combining information from multiple surveys through the empirical likelihood method," *The Canadian Journal of Statistics*, **32**, 15–26.