Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2010)
72, *Part* 1, *pp.* 27–48

Combining information from multiple surveys by using regression for efficient small domain estimation

Takis Merkouris

Athens University of Economics and Business, Greece

[Received April 2008. Final revision July 2009]

Summary. 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. Demand for small domain estimates has been growing in recent years among users of survey data. We explore the possibility of enhancing the precision of domain estimators by combining comparable information collected in multiple surveys of the same population. For this, 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 analytic results and an empirical study 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 design-based direct estimators proposed involve only domain-specific data on the variables of interest. This is in contrast with small domain (mostly small area) indirect estimators, based on a single survey, which incorporate through modelling data that are external to the targeted small domains. The approach proposed is also highly effective in handling the closely related problem of estimation for rare population characteristics.

Keywords: Auxiliary information; Calibration; Composite estimator; Generalized regression estimator; Rare populations; Small area estimation

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 much lesser degree because of the extra variability that is 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.

Variants of the regression estimation technique, which incorporates auxiliary survey information in derived estimators, have been introduced recently to enhance the reliability of domain estimators; see Estevao and Särndal (1999, 2004), Hidiroglou and Patak (2004) and references therein. The primary considerations in these works relate to the amount of the auxiliary information and the level (population level, domain level or some intermediate level) for which auxiliary variable totals are available, as well as to the type of regression estimator that is used. Briefly put, intuitive arguments and some formal evidence suggest that it may be advantageous to use

Address for correspondence: Takis Merkouris, Department of Statistics, Athens University of Economics and Business, 76 Patision Street, Athens 10434, Greece.
E-mail: merkouris@aueb.gr

 1369–7412/10/72027

domain-specific auxiliary information. The issues that may arise in the use of such auxiliary information in domain estimation are

(a) auxiliary variable totals at the domain level may not be available or may not be sufficient for enhancing reliability or may not be themselves of acceptable quality and
(b) the domain sample size may be too small to maintain the desirable properties of regression estimation.

The latter issue relates to the notion of a small domain, i.e. any domain with sample size too small to yield estimates of adequate precision even with the use of auxiliary information. Since the 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 sizable 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 which are external to the small areas of interest or are from other time periods. The domain estimators derived are then *indirect* estimators, in the sense that they incorporate data on the variables of interest which 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) and Rao (2003), page 23) but there seems to be a paucity of related research up to now. Yet, diverse possibilities in modern survey sampling practice abound. An array of relevant examples includes the following.

(a) Two or more separate surveys of the same population having some target variables in common: examples include various income surveys or consumer expenditure surveys that are conducted by many national statistical agencies, such as the Survey of Labour and Income Dynamics, Survey of Household Spending and Survey of Financial Security of Statistics Canada, and the 'diary' and 'interview' components of the Consumer Expenditure Survey of the US Bureau of Labor Statistics.
(b) One survey comprising two or more independent samples: these can be distinguished by
    (i) partial overlap in target variables, as in surveys with a split-questionnaire sampling design (see Raghunathan and Grizzle (1995)) and
    (ii) additional target variables in some of the samples (e.g. *structural* variables in some of the subsamples constituting Labour Force Surveys in the European Union).
(c) Two or more separate surveys of the same population, or one survey comprising two or more independent samples, with some auxiliary variables in common for which no known totals are available: this includes the setting of a large-scale survey with 'supplements', and the setting of *non-nested* double sampling (Hidiroglou, 2001).
(d) One main and one supplementary survey, either with the same target variables or with fewer target variables in the supplement, with the supplementary survey having the specific objective of improving the efficiency of small domain estimation: two important cases involve
    (i) an independent supplementary sample from the same entire frame, for targeting small domains cutting across strata (this sampling scheme may also be effective in

surveying rare populations, e.g. population of rare crops, or population of aboriginal people) and

(ii) an independent supplementary sample from the frames of strata containing or coinciding with the domains of interest (typically small areas).

It is to be noted that harmonization of the overlapping survey information is inbuilt in settings (b) and (d) above, whereas it may or may not occur in settings (a) and (c). In practice, common questionnaire items between the various surveys may be estimating different population quantities owing to discrepancies between definitions, survey frames and non-sampling errors caused by a difference in survey mode and non-response pattern. This complication can be rectified if the survey processes involved are harmonized at the stage of survey design. Harmonization of concepts, methodology and processes ensuring comparability of estimates for common items for surveys of similar content (e.g. income surveys) is feasible within the same organization. This is increasingly happening among national statistical agencies; for examples of such harmonization designed into similar surveys, see Statistics Netherlands (1998) and Webber *et al.* (2000). Such estimates may then be combined into more efficient composite estimates. The framework of the composite estimation methodology proposed encompasses all the above multiple-survey settings assuming that comparability of common survey items is ensured essentially by design; a modification of the methodology addressing departure from this assumption is included in Section 5.

Combining information from multiple surveys for more precise estimation of survey characteristics at the population level has been the subject of recent research by Renssen and Nieuwenbroek (1997) and Merkouris (2004), who used variants of generalized regression, and by Wu (2004), who took an empirical likelihood approach. Combining information at domain level was dealt with by Zieschang (1990), who used an antecedent of the aforementioned regression procedures with primary motivation to align estimates between the 'diary' and 'interview' components of the US Consumer Expenditure Survey for a variety of domains. Two recent references on small area estimation, by Elliott and Davis (2005) and Raghunathan *et al.* (2007), described two approaches for combining information from two US surveys to construct small area estimates of cancer risk factor prevalence. The first approach uses a propensity score extension of dual frame estimation, whereas the second uses a hierarchical Bayesian model; in both approaches, the objective is to address non-response and non-coverage errors in the two surveys and to provide compromise estimates.

In this paper, an adaptation of the regression procedure of Merkouris (2004) is used in small domain estimation, in conjunction with various options for incorporating auxiliary survey information in the estimation process. The regression method proposed 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 based, and the resulting domain estimators are direct because they involve only domain-specific data on the variables of interest. In particular, this design-based approach greatly enhances the precision 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 estimation method proposed is equally suitable for small geographic and non-geographic domains. It is also especially effective in handling the closely related problem of estimating rare population characteristics.

The organization of the paper is as follows. In Section 2, after setting out the notation and terminology, three alternative domain estimators derived from variants of the generalized regression (GREG) estimation procedure are assessed in terms of their relative efficiency. In Section 3,

an extension of each of the procedures of Section 2 is used to combine information from two surveys. The relative efficiency of these procedures is assessed analytically under certain conditions. In Section 4, an empirical study based on a complex survey is presented. In the concluding Section 5, theoretical and practical aspects of the estimation method proposed are discussed, including a modification of it devised to handle possible differences between combined totals.

2. Regression estimation for domains

2.1. Basic notation and terminology

Let $U = \{1, \ldots, k, \ldots, N\}$ denote a finite population, from which a probability sample s of size n is drawn according to a sampling design with known first- and second-order inclusion probabilities π_k and π_{kl} $(k, l \in U)$. Consider the sampling weight vector $\mathbf{w}$ with kth entry defined as $w_k = (1/\pi_k)\, I(k \in s)$, where I denotes the indicator variable, and let $\mathbf{Y}$ denote the $N \times r$ population matrix of an r-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 $\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{w}$ $(= \Sigma_U w_k y_k)$. For the $N \times p$ population matrix $\mathbf{X}$ of a p-dimensional auxiliary variable $\mathbf{x}$, assume that the total $\mathbf{t_x} = \mathbf{X}'\mathbf{1}$ is known. Let also $\mathbf{\Lambda}$ be the diagonal 'weighting' matrix that has w_k/q_k as kth entry, where q_k is a positive constant—the typical default value $q_k = 1$ for all k will be assumed, unless indicated otherwise. The subvectors and submatrices corresponding to the sample are designated by s. An n-dimensional vector of 'calibrated' weights, $\mathbf{c}_s$, 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

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

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

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

which has the equivalent form of a GREG estimator

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

where $\hat{\mathbf{X}} = \mathbf{X}_s'\mathbf{w}_s$ is the HT estimator of $\mathbf{t_x}$, and $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{Y}_s$ is the matrix of sample regression coefficients. The term $(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})'\mathbf{w}_s$ in estimator (3) is the sum of weighted sample regression residuals. By construction the GREG estimator (3) has the calibration property that $\hat{\mathbf{X}}^{\mathrm{R}} = \mathbf{t_x}$, i.e. 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 was given in Deville and Särndal (1992), and an extensive discussion of it is given in Särndal *et al.* (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 $\mathbf{Y}$ with the entries of the kth row set equal to 0 if $k \notin U_d$; accordingly, $\mathbf{Y}_{\bar{d}}$ denotes the matrix $\mathbf{Y}$ with the entries of the kth row set equal to 0 if $k \notin U_{\bar{d}}$. We can write then $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{Y}_d + \mathbf{Y}_{\bar{d}}$, and similarly for the matrix $\mathbf{X}$. Assuming that membership in U_d for every sample unit is observed, we denote by $\mathbf{Y}_{s_d}$ and $\mathbf{X}_{s_d}$ the associated sample domain quantities. The HT estimator of the domain total $\mathbf{t}_{\mathbf{y}_d} = \mathbf{Y}_d'\mathbf{1}$ is $\hat{\mathbf{Y}}_d = \mathbf{Y}_{s_d}'\mathbf{w}_s$.

2.2. Three regression estimators for domains

Three regression domain estimators of $\mathbf{t}_{\mathbf{y}_d}$ differing in the use of auxiliary information are presented below.

2.2.1. Regression at the population level

For any domain U_d, the vector of calibrated weights given by equation (1) can be used to generate, by $\mathbf{Y}'_{s_d}\mathbf{c}_s$, the GREG domain estimator of $\mathbf{t}_{\mathbf{y}_d}$

$$\hat{\mathbf{Y}}_d^{\mathrm{R}} = \hat{\mathbf{Y}}_d + \hat{\boldsymbol{\beta}}'_d(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{X}}), \tag{4}$$

where $\hat{\boldsymbol{\beta}}_d = (\mathbf{X}'_s\boldsymbol{\Lambda}_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\boldsymbol{\Lambda}_s\mathbf{Y}_{s_d}$. The vector of population residuals corresponding to $\hat{\mathbf{Y}}_d^{\mathrm{R}}$ is $\mathbf{E}_d = \mathbf{Y}_d - \mathbf{X}\boldsymbol{\beta}_d$, where $\boldsymbol{\beta}_d = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_d$. A useful alternative expression for $\mathbf{E}_d$ that will be recalled repeatedly below is $\mathbf{E}_d = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}_d$, with $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Using a matrix formulation of a standard result (see, for example, Särndal *et al.* (1992), page 235), the approximate design variance of $\hat{\mathbf{Y}}_d^{\mathrm{R}}$, which is denoted by $\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})$, is given by

$$\begin{aligned}\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}}) &= \mathbf{E}'_d\boldsymbol{\Lambda}^0\mathbf{E}_d \\ &= (\mathbf{Y}_d - \mathbf{X}\boldsymbol{\beta}_d)'\boldsymbol{\Lambda}^0(\mathbf{Y}_d - \mathbf{X}\boldsymbol{\beta}_d) \\ &= \mathbf{Y}'_d(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}_d,\end{aligned} \tag{5}$$

where $\boldsymbol{\Lambda}^0$ is a non-negative definite matrix whose klth entry is $(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l$, $\pi_{kk} \equiv \pi_k$. This formula assumes no adjustment of sampling weights for non-response. Noting that $\mathbf{X}'\boldsymbol{\Lambda}^0\mathbf{X}$, $\mathbf{Y}'_d\boldsymbol{\Lambda}^0\mathbf{Y}_d$ and $\mathbf{Y}'_d\boldsymbol{\Lambda}^0\mathbf{X}$ are the variance matrices $V(\hat{\mathbf{X}})$ and $V(\hat{\mathbf{Y}}_d)$ and the covariance matrix $\mathrm{cov}(\hat{\mathbf{Y}}_d, \hat{\mathbf{X}})$ respectively, equation (5) can be expressed as

$$\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}}) = V(\hat{\mathbf{Y}}_d) - \boldsymbol{\beta}'_d(\mathrm{cov}(\hat{\mathbf{Y}}_d, \hat{\mathbf{X}}))' - \mathrm{cov}(\hat{\mathbf{Y}}_d, \hat{\mathbf{X}})\boldsymbol{\beta}_d + \boldsymbol{\beta}'_d\, V(\hat{\mathbf{X}})\boldsymbol{\beta}_d. \tag{6}$$

Depending on the strength of correlation between $\mathbf{y}$ and $\mathbf{x}$, the use of the auxiliary total $\mathbf{t}_{\mathbf{x}}$ in the GREG estimation procedure may yield a highly efficient GREG estimator (3). This, however, may not hold for the GREG domain estimator (4); it may not even be true that estimator (4) is more efficient than the HT estimator $\hat{\mathbf{Y}}_d$. For simple random sampling and under certain simplifying assumptions, Estevao and Särndal (1999) showed that the *ratio* domain estimator (a special case of estimator (4)) may have only marginally smaller variance than the HT estimator, even if the domain is as large as half the population and the correlation between scalars y and x within the domain is high. Further argument is provided by the following consideration. In the common situation involving a single binary variable y and an $\mathbf{x}$ such that $\mathbf{1}_s = \mathbf{X}_s\mathbf{h}$ for a p-vector $\mathbf{h}$, strong correlation between y and $\mathbf{x}$ entails $y_k = 1 = \mathbf{x}_k\mathbf{h}$ for a large number of population units; thus, it also entails a large number of approximately zero regression residuals. Now, since the entries of the vector $\mathbf{Y}_d$ are defined to be 0 if they correspond to units outside the domain U_d, the linear relationship between y and $\mathbf{x}$ diminishes as the size of U_d decreases. It can be shown then that the number of standardized (by the fitted values for the y_ks) residuals which are close to 0 decreases also, assuming constant correlation between y and $\mathbf{x}$ in different domains, and thus the beneficial effect of regression at the population level is not sustained in domain estimation. Incidentally, this argument reveals also that this regression procedure is not effective in reducing the variability of estimators for variables that represent rare population characteristics (i.e. when U_d denotes such a population and $y_k = 1$ for $k \in U_d$, and $y_k = 0$ for $k \notin U_d$), regardless of the level at which they are computed. The foregoing considerations point to the need to use the auxiliary information at the domain level.

2.2.2. Regression at the domain level

Using the domain-specific quantities $\mathbf{X}_{s_d}$, $\mathbf{t}_{\mathbf{x}_d}$ and $\mathbf{Y}_{s_d}$ in equations (1) and (2) we obtain

$$\check{\mathbf{Y}}_d^{\mathrm{R}} = \hat{\mathbf{Y}}_d + \check{\boldsymbol{\beta}}'_d(\mathbf{t}_{\mathbf{x}_d} - \hat{\mathbf{X}}_d), \tag{7}$$

where $\check{\boldsymbol{\beta}}_d = (\mathbf{X}'_{s_d}\boldsymbol{\Lambda}_{s_d}\mathbf{X}_{s_d})^{-1}\mathbf{X}'_{s_d}\boldsymbol{\Lambda}_{s_d}\mathbf{Y}_{s_d}$. The corresponding population vector of residuals is $\mathbf{E}_d = \mathbf{Y}_d - \mathbf{X}_d\boldsymbol{\beta}_d$, with $\boldsymbol{\beta}_d = (\mathbf{X}'_d\mathbf{X}_d)^{-1}\mathbf{X}'_d\mathbf{Y}_d$. Writing $\mathbf{E}_d = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_d})\mathbf{Y}_d$, with $\mathbf{P}_{\mathbf{X}_d} = \mathbf{X}_d(\mathbf{X}'_d\mathbf{X}_d)^{-1}\mathbf{X}'_d$, we obtain

$$\begin{aligned}\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}}) &= \mathbf{Y}'_d(\mathbf{I} - \mathbf{P}_{\mathbf{X}_d})\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{X}_d})\mathbf{Y}_d \\ &= V(\hat{\mathbf{Y}}_d) - \boldsymbol{\beta}'_d(\mathrm{cov}(\hat{\mathbf{Y}}_d, \hat{\mathbf{X}}_d))' - \mathrm{cov}(\hat{\mathbf{Y}}_d, \hat{\mathbf{X}}_d)\boldsymbol{\beta}_d + \boldsymbol{\beta}'_d V(\hat{\mathbf{X}}_d)\boldsymbol{\beta}_d. \end{aligned} \tag{8}$$

It should be noted that the domain total $\mathbf{t}_{\mathbf{x}_d}$ that is used in expression (7) may not be readily available or may be of questionable quality, especially for very small domains. Moreover, the domain sample size may be very small or the number of auxiliary variables may be too large for the available domain sample size; this can cause significant bias in $\check{\mathbf{Y}}_d^{\mathrm{R}}$, and inflation of its variance due to the instability of $\check{\boldsymbol{\beta}}_d$. These potential weaknesses aside, the present regression set-up sets the limit regarding the extent to which auxiliary information can be exploited for estimation of the domain total $\mathbf{t}_{\mathbf{y}_d}$.

If we use $\mathbf{Y}_s\ (= \mathbf{Y}_{s_d} + \mathbf{Y}_{s_{\bar{d}}})$ in equation (2), while keeping $\mathbf{X}_{s_d}$ in place of $\mathbf{X}_s$, we obtain the population level estimator $\check{\mathbf{Y}}_d^{\mathrm{R}} + \hat{\mathbf{Y}}_{\bar{d}}$ of $\mathbf{t}_{\mathbf{y}_d}$, where $\hat{\mathbf{Y}}_{\bar{d}}$ is the HT estimator of $\mathbf{t}_{\mathbf{y}_{\bar{d}}}$. Domain estimators for both domains U_d and $U_{\bar{d}}$ can be obtained simultaneously by using the augmented matrix $(\mathbf{X}_{s_d} \quad \mathbf{X}_{s_{\bar{d}}})$ and the corresponding vector of totals $(\mathbf{t}'_{\mathbf{x}_d}, \mathbf{t}'_{\mathbf{x}_{\bar{d}}})'$ in equation (1); the matrices $\mathbf{X}_{s_d}$ and $\mathbf{X}_{s_{\bar{d}}}$ need not have any columns corresponding to the same auxiliary variables. Then equation (2) yields the GREG estimator of $\mathbf{t}_{\mathbf{y}}$ as $\check{\mathbf{Y}}^{\mathrm{R}} = \check{\mathbf{Y}}_d^{\mathrm{R}} + \check{\mathbf{Y}}_{\bar{d}}^{\mathrm{R}}$. This can be generalized to any number of mutually exclusive and exhaustive domains of U.

2.2.3. *Regression incorporating the domain size*

A more sensible domain estimator than estimator (7) may involve the population level auxiliary total $\mathbf{t}_{\mathbf{x}}$, and as additional domain level auxiliary information only the domain size N_d. Such an estimator may be obtained by a mixture of the two regression procedures that were described above. Formally, for the augmented matrix $\boldsymbol{\mathcal{X}}_s = (\mathbf{X}_s\ \mathbf{1}_{s_d})$ and the corresponding vector of totals $\mathbf{t}_{\boldsymbol{\mathcal{X}}} = (\mathbf{t}'_{\mathbf{x}}, N_d)'$, a useful decomposition of the vector of calibrated weights $\mathbf{c}_s = \mathbf{w}_s + \boldsymbol{\Lambda}_s\boldsymbol{\mathcal{X}}_s(\boldsymbol{\mathcal{X}}'_s\boldsymbol{\Lambda}_s\boldsymbol{\mathcal{X}}_s)^{-1}(\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'_s\mathbf{w}_s)$ is obtained as

$$\mathbf{c}_s = \mathbf{c}_{xs} + \mathbf{L}_s\mathbf{1}_{s_d}(\mathbf{1}'_{s_d}\mathbf{L}_s\mathbf{1}_{s_d})^{-1}(N_d - \mathbf{1}'_{s_d}\mathbf{c}_{xs}), \tag{9}$$

where $\mathbf{c}_{xs} = \mathbf{w}_s + \boldsymbol{\Lambda}_s\mathbf{X}_s(\mathbf{X}'_s\boldsymbol{\Lambda}_s\mathbf{X}_s)^{-1}(\mathbf{t}_{\mathbf{x}} - \mathbf{X}'_s\mathbf{w}_s)$, $\mathbf{L}_s = \boldsymbol{\Lambda}_s(\mathbf{I} - \mathbf{P}_{\mathbf{X}_s})$, $\mathbf{P}_{\mathbf{X}_s} = \mathbf{X}_s(\mathbf{X}'_s\boldsymbol{\Lambda}_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\boldsymbol{\Lambda}_s$; for the decomposition of $\mathbf{c}_s$ in a more general set-up, see Renssen and Nieuwenbroek (1997) and Merkouris (2004). Then the domain estimator of $\mathbf{t}_{\mathbf{y}_d}$, which is denoted by $\tilde{\mathbf{Y}}_d^{\mathrm{R}}$, corresponding to the partitioned regression matrix $\boldsymbol{\mathcal{X}}_s$, takes the form

$$\tilde{\mathbf{Y}}_d^{\mathrm{R}} = \hat{\mathbf{Y}}_d^{\mathrm{R}} + \tilde{\boldsymbol{\beta}}'_d(N_d - \hat{N}_d^{\mathrm{R}}), \tag{10}$$

where $\tilde{\boldsymbol{\beta}}_d = (\mathbf{1}'_{s_d}\mathbf{L}_s\mathbf{1}_{s_d})^{-1}\mathbf{1}'_{s_d}\mathbf{L}_s\mathbf{Y}_{s_d}$. It is not difficult to verify that the population vector of residuals for the estimator (10) is $\mathbf{E}_d = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\mathbf{Y}_d - \mathbf{1}_d\boldsymbol{\beta}_d)$, with $\boldsymbol{\beta}_d = (\mathbf{1}'_d\mathbf{L}\mathbf{1}_d)^{-1}\mathbf{1}'_d\mathbf{L}\mathbf{Y}_d$, $\mathbf{L} = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$ and $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. In view of equation (5), the approximate design variance of $\tilde{\mathbf{Y}}_d^{\mathrm{R}}$ is given by

$$\begin{aligned}\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}}) &= (\mathbf{Y}_d - \mathbf{1}_d\boldsymbol{\beta}_d)'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{X}})(\mathbf{Y}_d - \mathbf{1}_d\boldsymbol{\beta}_d) \\ &= \mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}}) - \boldsymbol{\beta}'_d(\mathrm{AC}(\hat{\mathbf{Y}}_d^{\mathrm{R}}, \hat{N}_d^{\mathrm{R}}))' - \mathrm{AC}(\hat{\mathbf{Y}}_d^{\mathrm{R}}, \hat{N}_d^{\mathrm{R}})\boldsymbol{\beta}_d + \boldsymbol{\beta}'_d\,\mathrm{AV}(\hat{N}_d^{\mathrm{R}})\boldsymbol{\beta}_d, \end{aligned} \tag{11}$$

where AC denotes approximate covariance. For calibration to both domain sizes N_d and $N_{\bar{d}}$, the regression procedure requires the matrix $(\mathbf{X}_s\ \mathbf{1}_{s_d}\ \mathbf{1}_{s_{\bar{d}}})$ and the vector of totals $\mathbf{t}_{\boldsymbol{\mathcal{X}}} = (\mathbf{t}'_{\mathbf{x}}, N_d, N_{\bar{d}})'$,

unless N can be obtained by adding up components of $\mathbf{t_x}$ (and calibration to $N_{\bar{d}}$ is then automatic). Estimators of $\mathbf{t}_{\mathbf{y}_d}$ and $\mathbf{t}_{\mathbf{y}_{\bar{d}}}$ will be obtained then from the modification of equation (9) $\mathbf{c}_s = \mathbf{c}_{xs} + \mathbf{L}_s \mathbf{1}_{s_{d\bar{d}}} (\mathbf{1}'_{s_{d\bar{d}}} \mathbf{L}_s \mathbf{1}_{s_{d\bar{d}}})^{-1} (\mathbf{N} - \mathbf{1}'_{s_{d\bar{d}}} \mathbf{c}_{xs})$, where $\mathbf{1}_{s_{d\bar{d}}}$ denotes the submatrix $(\mathbf{1}_{s_d} \mathbf{1}_{s_{\bar{d}}})$ and $\mathbf{N}$ denotes the vector of domain sizes $(N_d, N_{\bar{d}})'$. Obviously, this can be generalized to include the sizes of any number of mutually exclusive and exhaustive domains of U. Unlike the estimator (7), the resulting estimator for a specific domain will depend on the other domains that are involved in the regression procedure.

The incorporation of domain sizes into a regression procedure for efficient domain estimation was suggested also by Estevao and Särndal (1999). Their regression procedure involves a single variable x and a number of domains forming a partition of U but is otherwise equivalent to that used here. However, their estimator is expressed in terms of GREG (*ratio*) estimators of $\mathbf{t}_{\mathbf{y}_d}$ and $\mathbf{t_x}$ involving calibration to the domain sizes, whereas estimator (10) is expressed in terms of GREG estimators of $\mathbf{t}_{\mathbf{y}_d}$ and N_d involving calibration to the vector of totals $\mathbf{t_x}$. The form (10) and its extension to several domains are novel and allow direct comparison of $\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})$ with $\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})$ for each domain.

2.2.4. *Comparing the three domain regression estimators*

The estimators (4), (7) and (10) are compared with respect to their approximate design variances under the conditions of the following theorem; the proof is given in Appendix A.

Theorem 1.

(a) Suppose that $\mathbf{1} = \mathbf{Xh}$, for a constant p-vector $\mathbf{h}$. Then, under simple random sampling without replacement the following inequalities hold.

$$\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}}) \leqslant \mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}}) \leqslant \mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}}) \leqslant V(\hat{\mathbf{Y}}_d). \tag{12}$$

(b) Assume stratified simple random sampling without replacement and combined regression with an intercept for each stratum, i.e., with auxiliary row vector corresponding to unit k, $\mathbf{x}'_k = (\mathbf{x}'_{0k}, d_{1k}, \ldots, d_{Hk})$, where $\mathbf{x}'_{0k}$ denotes the row p-vector of the core auxiliary variables and d_{lk} is the indicator variable of the membership of population unit k to stratum $l = 1, \ldots, H$. Then, if the constants q_k in $\mathbf{\Lambda}_s$ are specified as $q_k = n_l(N_l - 1)/N_l^2(1 - f_l)$ for all units of stratum l, f_l denoting the sampling fraction n_l/N_l for stratum l, the inequalities in expression (12) hold.

(c) Assume Poisson sampling, and specify the constants q_k in $\mathbf{\Lambda}_s$ as $q_k = \pi_k/(1 - \pi_k)$. Then the inequality $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}}) \leqslant \mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})$ holds under the condition $\mathbf{1} = \mathbf{Xh}$. All other inequalities in expression (12) hold without this condition.

Expression (12) shows that under the conditions of theorem 1 all three regression estimators of the domain total $\mathbf{t}_{\mathbf{y}_d}$ are more efficient than the HT estimator, with more domain-specific auxiliary information leading to more efficient estimation of $\mathbf{t}_{\mathbf{y}_d}$. This inequality holds in the partial ordering of non-negative definite matrices, and therefore it also holds for any linear combination of the components of each of the estimators involved. The condition $\mathbf{1} = \mathbf{Xh}$ in theorem 1 is customarily satisfied in surveys that use GREG estimation; it is obviously satisfied in the set-up of part (b). The inequality $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}}) \leqslant \mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})$ in (a) was proved also, in a different way, for a single variable y and as a strict inequality, by Hidiroglou and Patak (2004) on the assumption that an intercept is included in the regression, which is a special case of the condition $\mathbf{1} = \mathbf{Xh}$. The inequality $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}}) \leqslant \mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})$ in (c) was proved also, in a different way and for a single variable y, by Estevao and Särndal (1999). The inequality $\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}}) \leqslant V(\hat{\mathbf{Y}}_d)$ in (b) and (c)

was proved also, in a different way, for a single variable y and at population level, by Särndal (1996).

In each of parts (b) and (c) of theorem 1 the specification of q_k implies the modified matrices $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}$ and $\mathbf{P}_{\mathbf{X}_d} = \mathbf{X}_d(\mathbf{X}_d'\mathbf{Q}\mathbf{X}_d)^{-1}\mathbf{X}_d'\mathbf{Q}$, where $Q = \text{diag}(q_1^{-1}, \ldots, q_N^{-1})$. When the inclusion probabilities are equal for all units, the results hold for the default value $q_k = 1$ for all k, provided that the ratio $N_l/(N_l - 1)$ is approximately equal to 1 in the case of stratified simple random sampling without replacement. The extension of part (c) to stratified Poisson sampling is trivial.

Although analytic results are confined to the above designs, foregoing intuitive arguments suggest that these results would hold under more general designs.

3. Domain estimation using information from two surveys

We assume that there are samples s_1 and s_2 of sizes n_1 and n_2 respectively from two separate surveys of the same target population, a vector $\mathbf{z}$ of q survey variables that are 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. We also assume that domains of interest are identifiable in the data files of both surveys. The present framework may include one survey comprising two independent samples, thus encompassing all types of comparable pairs of samples that were listed in Section 1. 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 by using special regression set-ups for the combined sample as follows.

3.1. Regression at the U-level and combining information at the U_d-level

The set-up of simultaneous regression for the two samples involving the block diagonal matrices $\mathbf{X}_s = \text{diag}(\mathbf{X}_{s_i})$ and $\mathbf{\Lambda}_s = \text{diag}(\mathbf{\Lambda}_{s_i})$, and the vectors $\mathbf{w}_s = (\mathbf{w}_{s_1}', \mathbf{w}_{s_2}')'$ and $\mathbf{t} = (\mathbf{t}_{\mathbf{x}_1}', \mathbf{t}_{\mathbf{x}_2}')'$, generates a vector of calibrated weights $\mathbf{c}_{xs}$ that is given by

$$\mathbf{c}_{xs} = \begin{pmatrix} \mathbf{w}_{s_1} \\ \mathbf{w}_{s_2} \end{pmatrix} + \begin{pmatrix} \mathbf{\Lambda}_{s_1}\mathbf{X}_{s_1}(\mathbf{X}_{s_1}'\mathbf{\Lambda}_{s_1}\mathbf{X}_{s_1})^{-1}(\mathbf{t}_{\mathbf{x}_1} - \mathbf{X}_{s_1}'\mathbf{w}_{s_1}) \\ \mathbf{\Lambda}_{s_2}\mathbf{X}_{s_2}(\mathbf{X}_{s_2}'\mathbf{\Lambda}_{s_2}\mathbf{X}_{s_2})^{-1}(\mathbf{t}_{\mathbf{x}_2} - \mathbf{X}_{s_2}'\mathbf{w}_{s_2}) \end{pmatrix}. \tag{13}$$

The two components of $\mathbf{c}_{xs}$ give rise to two independent GREG domain estimators of the domain total $\mathbf{t}_{\mathbf{z}_d}$ of the type shown in expression (4). Combining information on $\mathbf{z}$ at the level of domain U_d is accomplished by incorporating in the regression procedure the additional calibration constraint that the two estimators of $\mathbf{t}_{\mathbf{z}_d}$ are calibrated to each other, i.e. they are aligned. This involves the extended regression matrix and the corresponding vector of control totals

$$\mathcal{X}_s = \begin{pmatrix} \mathbf{X}_{s_1} & \mathbf{0} & \mathbf{Z}_{s_1 d} \\ \mathbf{0} & \mathbf{X}_{s_2} & -\mathbf{Z}_{s_2 d} \end{pmatrix}, \qquad \mathbf{t} = \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1} \\ \mathbf{t}_{\mathbf{x}_2} \\ \mathbf{0} \end{pmatrix}. \tag{14}$$

Now assume that $(\mathbf{X}_{s_i} \quad \mathbf{Z}_{s_{id}})$, $i = 1, 2$, is of full rank $p_i + q$ and write $\mathcal{X}_s$ in partition form as $\mathcal{X}_s = (\mathbf{X}_s \quad \mathcal{Z}_{s_d})$, where $\mathbf{X}_s$ and $\mathcal{Z}_{s_d}$ are of dimension $(n_1 + n_2) \times (p_1 + p_2)$ and $(n_1 + n_2) \times q$ respectively. For samples s_1 and s_2 with arbitrary sampling designs, reset the default value $q_{ik} = 1$ in the entries of $\mathbf{\Lambda}_{s_i}$ to $q_{ik} = \tilde{n}_i$ for every unit k of survey i, where $\tilde{n}_i = n_i/d_i$ are the effective sample sizes—d_i denoting design effects. 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 $\mathbf{X}_s = \text{diag}(\mathbf{X}_{s_i})$ implies that $\mathbf{L}_s = \text{diag}(\mathbf{L}_{s_i})$, 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 $\mathbf{\Lambda}_s = \text{diag}(\mathbf{\Lambda}_{s_i})$, the regression procedure based on the partitioned matrix $\mathcal{X}_s$ generates the vector of calibrated weights

$$\mathbf{c}_s = \mathbf{c}_{xs} + \mathbf{L}_s \mathcal{Z}_{s_d} (\mathcal{Z}'_{s_d} \mathbf{L}_s \mathcal{Z}_{s_d})^{-1} (\mathbf{0} - \mathcal{Z}'_{s_d} \mathbf{c}_{xs})$$
$$= \begin{pmatrix} \mathbf{c}_{xs_1} \\ \mathbf{c}_{xs_2} \end{pmatrix} + \begin{pmatrix} \mathbf{L}_{s_1} \mathbf{Z}_{s_{1d}} \\ -\mathbf{L}_{s_2} \mathbf{Z}_{s_{2d}} \end{pmatrix} (\mathbf{Z}'_{s_{1d}} \mathbf{L}_{s_1} \mathbf{Z}_{s_{1d}} + \mathbf{Z}'_{s_{2d}} \mathbf{L}_{s_2} \mathbf{Z}_{s_{2d}})^{-1} (\mathbf{Z}'_{s_{2d}} \mathbf{c}_{xs_2} - \mathbf{Z}'_{s_{1d}} \mathbf{c}_{xs_1}).$$

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

$$\begin{aligned} \hat{Y}^{\mathrm{CR}}_{1d} &= \hat{Y}^{\mathrm{R}}_{1d} + \hat{\mathbf{B}}'_{y_{1d}} (\mathbf{I} - \hat{\mathbf{B}}_d)(\hat{\mathbf{Z}}^{\mathrm{R}}_{2d} - \hat{\mathbf{Z}}^{\mathrm{R}}_{1d}), \\ \hat{Y}^{\mathrm{CR}}_{2d} &= \hat{Y}^{\mathrm{R}}_{2d} - \hat{\mathbf{B}}'_{y_{2d}} \hat{\mathbf{B}}_d (\hat{\mathbf{Z}}^{\mathrm{R}}_{2d} - \hat{\mathbf{Z}}^{\mathrm{R}}_{1d}), \end{aligned} \tag{15}$$

where $\hat{\mathbf{B}}_{y_{id}} = (\mathbf{Z}'_{s_{id}} \mathbf{L}_{s_i} \mathbf{Z}_{s_{id}})^{-1} \mathbf{Z}'_{s_{id}} \mathbf{L}_{s_i} \mathbf{Y}_{s_{id}}$, $\hat{\mathbf{B}}_d = \mathbf{Z}'_{s_{2d}} \mathbf{L}_{s_2} \mathbf{Z}_{s_{2d}} (\mathbf{Z}'_{s_{1d}} \mathbf{L}_{s_1} \mathbf{Z}_{s_{1d}} + \mathbf{Z}'_{s_{2d}} \mathbf{L}_{s_2} \mathbf{Z}_{s_{2d}})^{-1}$ and $\hat{Y}^{\mathrm{R}}_{id}$ and $\hat{\mathbf{Z}}^{\mathrm{R}}_{id}$ are of the form shown in expression (4). For the q-dimensional common variable $\mathbf{z}$, we have the two identical (by construction) estimators of $\mathbf{t}_{\mathbf{z}_d}$

$$\begin{aligned} \hat{\mathbf{Z}}^{\mathrm{CR}}_{1d} &= \hat{\mathbf{Z}}^{\mathrm{R}}_{1d} + (\mathbf{I} - \hat{\mathbf{B}}_d)(\hat{\mathbf{Z}}^{\mathrm{R}}_{2d} - \hat{\mathbf{Z}}^{\mathrm{R}}_{1d}), \\ \hat{\mathbf{Z}}^{\mathrm{CR}}_{2d} &= \hat{\mathbf{Z}}^{\mathrm{R}}_{2d} - \hat{\mathbf{B}}_d (\hat{\mathbf{Z}}^{\mathrm{R}}_{2d} - \hat{\mathbf{Z}}^{\mathrm{R}}_{1d}), \end{aligned} \tag{16}$$

which can be written in the form of the composite GREG estimator

$$\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d} = \hat{\mathbf{Z}}^{\mathrm{CR}}_{2d} = \hat{\mathbf{B}}_d \hat{\mathbf{Z}}^{\mathrm{R}}_{1d} + (\mathbf{I} - \hat{\mathbf{B}}_d) \hat{\mathbf{Z}}^{\mathrm{R}}_{2d}. \tag{17}$$

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

$$\begin{aligned} \mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{1d}) = \mathrm{AV}(\hat{Y}^{\mathrm{R}}_{1d}) &+ \mathbf{B}'_{y_{1d}} (\mathbf{I} - \mathbf{B}_d) \{\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d}) + \mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{2d})\} (\mathbf{I} - \mathbf{B}_d)' \mathbf{B}_{y_{1d}} \\ &- 2\mathbf{B}'_{y_{1d}} (\mathbf{I} - \mathbf{B}_d)(\mathrm{AC}(\hat{Y}^{\mathrm{R}}_{1d}, \hat{\mathbf{Z}}^{\mathrm{R}}_{1d}))', \end{aligned} \tag{18}$$

where $\mathbf{B}_{y_{1d}} = (\mathbf{Z}'_d \mathbf{L}_1 \mathbf{Z}_d)^{-1} \mathbf{Z}'_d \mathbf{L}_1 \mathbf{Y}_d$ 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 $\mathbf{L}_i = (1/\tilde{n}_i)(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})$, are the population counterparts of $\hat{\mathbf{B}}_{y_{1d}}$ and $\hat{\mathbf{B}}_d$ respectively. Analogous is the expression of $\mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{2d})$. Furthermore, $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d})$ and $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{2d})$ are given by

$$\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d}) = \mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{2d}) = \mathbf{B}_d \, \mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d}) \mathbf{B}'_d + (\mathbf{I} - \mathbf{B}_d) \, \mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{2d})(\mathbf{I} - \mathbf{B}_d)'. \tag{19}$$

It is clear from the above formulation that estimates for common and non-common variables are obtained by using the data of only one of the surveys. Furthermore, 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 (15) and (17) which are more efficient than the regression estimators that are 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. With such reasoning, setting $q_{ik} = \tilde{n}_i$ in the entries of $\mathbf{\Lambda}_{s_i}$ entailed the weighting of the quadratic forms $\mathbf{Z}'_{s_{id}} \mathbf{\Lambda}_{s_i} (\mathbf{I} - \mathbf{P}_{\mathbf{X}_{s_i}}) \mathbf{Z}_{s_{id}}$ in inverse proportion to the $\tilde{n}_i$s, and as a result the coefficient $\hat{\mathbf{B}}_d$, which is generated implicitly by the regression procedure, accounts for the differential in effective sample size between the two samples. Values of q_{ik} that yield the most efficient composite estimators can be specified in certain situations (see relevant comments later in this section). An equivalent adjustment of the entries of $\mathbf{\Lambda}_{s_i}$, with the same effect on $\hat{\mathbf{B}}_d$, can be made through the scaling adjustment of the entries of $\mathbf{\Lambda}_{s_1}$ and $\mathbf{\Lambda}_{s_2}$ by $1 - \phi$ and ϕ respectively, where $\phi = \tilde{n}_1/(\tilde{n}_1 + \tilde{n}_2)$. A least squares characterization of $\hat{\mathbf{B}}_d$ and related efficiency considerations are as in Merkouris (2004).

It is important to note that without the above adjustment in the entries of $\mathbf{\Lambda}_{s_i}$ it would not necessarily follow that the composite estimators (15) and (17) would be more efficient than the regression estimators that are based on one sample. For instance, when the auxiliary variables that are used in the two surveys are the same, with the default setting $q_{ik}=1$ 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 $\mathbf{Z}'_d(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Z}_d$ and, therefore, the coefficients $\mathbf{B}_d$ and $\mathbf{I}-\mathbf{B}_d$ are both equal to $\frac{1}{2}\mathbf{I}$, so equation (19) becomes $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d})=\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{2d})=\frac{1}{4}\{\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})+\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{2d})\}$. It follows then that $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d})\leqslant\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})$ only if $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{2d})\leqslant 3\,\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})$. Such inefficient weighting of the variance components arises also, though less obviously, when the auxiliary vectors that are used in the two surveys are not identical. In the case of simple random sampling without replacement for both surveys with sampling fractions $f_i=n_i/N$, and with $\mathbf{1}=\mathbf{X}_i\mathbf{h}_i$, for constant p_i-vectors $\mathbf{h}_i$, it can be shown (as part of the proof of theorem 1) that $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{id})=\lambda^0_i\mathbf{Z}'_d(\mathbf{I}-\mathbf{P}_{\mathbf{X}})\mathbf{Z}_d$, where $\lambda^0_i=N^2(1-f_i)/n_i(N-1)$. This implies that $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d})\leqslant\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})$ only if $\lambda^0_2\leqslant 3\lambda^0_1$, and if the ratio of the finite population correction factors $1-f_1$ and $1-f_2$ is approximately equal to 1, $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d})\leqslant\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})$ only if $n_2\geqslant n_1/3$. In words, with $q_{ik}=1$ in $\mathbf{\Lambda}_{s_i}$ the composite GREG estimator $\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d}$ is more efficient than the GREG estimator $\hat{\mathbf{Z}}^{\mathrm{R}}_{1d}$ only if the size of sample s_2 is at least a third the size of s_1. As for the estimator $\hat{Y}^{\mathrm{CR}}_{1d}$, it can be easily verified that equation (18) reduces to

$$\mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{1d})=\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{1d})+\{(\lambda^0_2-3\lambda^0_1)/4\lambda^0_1\}\,\mathrm{AC}(\hat{Y}^{\mathrm{R}}_{1d},\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})\,\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})^{-1}(\mathrm{AC}(\hat{Y}^{\mathrm{R}}_{1d},\hat{\mathbf{Z}}^{\mathrm{R}}_{1d}))',$$

so $\mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{1d})\leqslant\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{1d})$ only if $\lambda^0_2\leqslant 3\lambda^0_1$ or if $n_2\geqslant n_1/3$. When $n_1=n_2$, $\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{1d})=\frac{1}{2}\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})$. Also,

$$\mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{1d})=\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{1d})-\tfrac{1}{2}\mathrm{AC}(\hat{Y}^{\mathrm{R}}_{1d},\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})\,\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})^{-1}\,\mathrm{AC}(\hat{Y}^{\mathrm{R}}_{1d},\hat{\mathbf{Z}}^{\mathrm{R}}_{1d})'.$$

By the Cauchy–Schwarz inequality, $\mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{1d})>\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{1d})-\frac{1}{2}\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{1d})$, and so $\frac{1}{2}\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{1d}<\mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{1d})\leqslant\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{1d})$.

Having shown analytically in Section 2.2 that the GREG domain estimator (4) is more efficient than the HT domain estimator under the conditions of theorem 1, it is interesting to see whether for the same sampling designs the composite GREG estimators (15) and (17) are more efficient than their single-sample components (which are of the type (4)). This is answered by the following proposition, the proof of which is given in Appendix A.

Proposition 1.

(a) Suppose that $\mathbf{1}=\mathbf{X}_i\mathbf{h}_i$, for constant p_i-vectors $\mathbf{h}_i$. For both surveys, assume simple random sampling without replacement with sampling fractions $f_i=n_i/N$ and specify the constants q_{ik} in $\mathbf{\Lambda}_{s_i}$ as $q_{ik}=n_i/(1-f_i)$. Then the following inequalities hold.

$$\begin{aligned}\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{CR}}_{id})&<\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{id}),\\ \mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{id})&<\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{id}).\end{aligned}\tag{20}$$

Furthermore, when $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the same variables and $(1-f_1)/(1-f_2)\approx 1$,

$$\begin{aligned}\mathrm{AV}\,(\hat{\mathbf{Z}}^{\mathrm{CR}}_{id})\mathrm{AV}(\hat{\mathbf{Z}}^{\mathrm{R}}_{id})^{-1}&=\{n_i/(n_1+n_2)\}\mathbf{I},\\ \mathrm{AV}(\hat{Y}^{\mathrm{CR}}_{id})\,\mathrm{AV}(\hat{Y}^{\mathrm{R}}_{id})^{-1}&>n_i/(n_1+n_2).\end{aligned}\tag{21}$$

(b) For each survey assume stratified simple random sampling without replacement and combined regression in which the auxiliary row vector corresponding to unit k for survey i is $\mathbf{x}'_{ik}=(\mathbf{x}'_{0ik},d_{1k},\ldots,d_{H_ik})$, where $\mathbf{x}'_{0i}$ denotes the row p_i-vector of the core auxiliary variables and d_{lk} is the indicator variable of the membership of population unit k to

stratum $l = 1, \ldots, H_i$. Then, if the constants q_{ik} in $\mathbf{\Lambda}_{s_i}$ are specified as $q_{ik} = n_{il}(N_{il} - 1)/N_{il}^2(1 - f_{il})$ for all units of stratum l, inequalities (20) hold. When the sampling fractions for all strata of survey i are equal to $f_i = n_i/N$ and $(1 - f_1)/(1 - f_2) \approx 1$, and the ratio $N_{il}/(N_{il} - 1)$ is approximately equal to 1, and in addition $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the same variables (including the same stratification for the two surveys), result (21) also holds.

(c) For both surveys assume Poisson sampling and specify the constants q_{ik} in $\mathbf{\Lambda}_{s_i}$ as $q_{ik} = \pi_{ik}/(1 - \pi_{ik})$. Then inequalities (20) hold. When the inclusion probabilities π_{ik} for survey i are all equal to $f_i = n_i/N$ and $(1 - f_1)/(1 - f_2) \approx 1$, and $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the same variables, result (21) also holds.

The equality in result (21) shows that under the stated conditions the efficiency of $\hat{\mathbf{Z}}_{id}^{\mathrm{CR}}$ relative to $\hat{\mathbf{Z}}_{id}^{\mathrm{R}}$ (componentwise) can be substantial (e.g. 50% reduction in variance if $n_1 = n_2$). For $\hat{Y}_{id}^{\mathrm{CR}}$, which borrows strength indirectly through the correlation of y_i with $\mathbf{z}$, the efficiency relative to $\hat{Y}_{id}^{\mathrm{R}}$ is lower—it is clear from equation (18) that the efficiency of $\hat{Y}_{id}^{\mathrm{CR}}$ depends on the strength of correlation between y_i and $\mathbf{z}$. Under the conditions of proposition 1, with the particular specifications of q_{ik}, it is shown in the proof that $\mathbf{B}'_{y_{1d}}(\mathbf{I} - \mathbf{B}_d) = \mathrm{AC}(\hat{Y}_{1d}^{\mathrm{R}}, \hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}$ and $\mathbf{B}_d = \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}$. These are the optimal (variance minimizing) coefficients in equations (18) and (19). In more general settings, the gain in efficiency will be somewhat smaller, as the coefficients $\mathbf{B}'_{y_{1d}}(\mathbf{I} - \mathbf{B}_d)$ and $\mathbf{B}_d$ (incorporating the generic specification $q_{ik} = \tilde{n}_i$) will only be approximations of the optimal cofficients. This is because, for general sampling designs, $\mathbf{B}_d$ may not precisely reflect the relative interaction of design and regression effects between the two surveys.

If we choose to combine information on a subset of the common variables, then for the rest we derive two domain estimators, as in expression (15), and it would then be 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 $\hat{Y}_d^{\mathrm{CR}} = \phi\hat{Y}_{1d}^{\mathrm{CR}} + (1 - \phi)\hat{Y}_{2d}^{\mathrm{CR}}$, where $\phi = \tilde{n}_1/(\tilde{n}_1 + \tilde{n}_2)$. In proposition 1, in particular, when equal inclusion probabilities for all units of survey i imply uniform q_i for survey i (ignoring the ratio $N_{il}/(N_{il} - 1)$ in (b)) the approximate variance of $\hat{Y}_d^{\mathrm{CR}}$ can be derived without difficulty as $\mathrm{AV}(\hat{Y}_d^{\mathrm{CR}}) = \phi^2\,\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{R}}) + (1 - \phi)^2\,\mathrm{AV}(\hat{Y}_{2d}^{\mathrm{R}}) - \mathbf{a}\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}\mathbf{a}'$, where $\mathbf{a} = \phi\,\mathrm{AC}(\hat{Y}_{1d}^{\mathrm{R}}, \hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) - (1 - \phi)\,\mathrm{AC}(\hat{Y}_{2d}^{\mathrm{R}}, \hat{\mathbf{Z}}_{2d}^{\mathrm{R}})$, and $\phi = q_1/(q_1 + q_2)$. Clearly, unless $\mathbf{x}_1$ and $\mathbf{x}_2$ represent the same variables (implying that $\mathbf{a} = \mathbf{0}$), the variance of $\hat{Y}_d^{\mathrm{CR}}$ is smaller than the variance of the composite $\phi\hat{Y}_{1d}^{\mathrm{R}} + (1 - \phi)\hat{Y}_{2d}^{\mathrm{R}}$ of the initial independent GREG estimators—maybe not by much, in view of $\mathbf{a}$. The computation of the composite $\hat{Y}_d^{\mathrm{CR}}$ can be incorporated in the GREG composite estimation procedure without difficulty, whereas an optimal linear combination of the dependent estimators $\hat{Y}_{1d}^{\mathrm{CR}}$ and $\hat{Y}_{2d}^{\mathrm{CR}}$ 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 set-up (14) 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 simply to indicate that for various domains the information on $\mathbf{z}$ from the two samples is combined at the domain level, and that expressions (15) and (17) give estimates for $\mathbf{t}_{y_{id}}$ and $\mathbf{t}_{\mathbf{z}_d}$ for each of these domains.

3.2. Regression and combining information at the U_d-level

With the rationale that was used in Section 2.2.2, we now introduce the variant of the regression set-up (14)

$$\mathcal{X}_{s_d} = \begin{pmatrix} \mathbf{X}_{s_{1d}} & \mathbf{0} & \mathbf{Z}_{s_{1d}} \\ \mathbf{0} & \mathbf{X}_{s_{2d}} & -\mathbf{Z}_{s_{2d}} \end{pmatrix}, \qquad \mathbf{t} = \begin{pmatrix} \mathbf{t}_{\mathbf{x}_{1d}} \\ \mathbf{t}_{\mathbf{x}_{2d}} \\ \mathbf{0} \end{pmatrix}, \tag{22}$$

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

$$\begin{aligned} \check{Y}^{\mathrm{CR}}_{1d} &= \check{Y}^{\mathrm{R}}_{1d} + \check{\mathbf{B}}'_{y_{1d}}(\mathbf{I} - \check{\mathbf{B}}_d)(\check{\mathbf{Z}}^{\mathrm{R}}_{2d} - \check{\mathbf{Z}}^{\mathrm{R}}_{1d}), \\ \check{Y}^{\mathrm{CR}}_{2d} &= \check{Y}^{\mathrm{R}}_{2d} - \check{\mathbf{B}}'_{y_{2d}}\check{\mathbf{B}}_d(\check{\mathbf{Z}}^{\mathrm{R}}_{2d} - \check{\mathbf{Z}}^{\mathrm{R}}_{1d}), \end{aligned} \tag{23}$$

and

$$\check{\mathbf{Z}}^{\mathrm{CR}}_{1d} = \check{\mathbf{Z}}^{\mathrm{CR}}_{2d} = \check{\mathbf{B}}_d\check{\mathbf{Z}}^{\mathrm{R}}_{1d} + (\mathbf{I} - \check{\mathbf{B}}_d)\check{\mathbf{Z}}^{\mathrm{R}}_{2d}, \tag{24}$$

where $\check{Y}^{\mathrm{R}}_{id}$ and $\check{\mathbf{Z}}^{\mathrm{R}}_{id}$ are GREG domain estimators of the type shown in equation (7), using auxiliary data only from U_d, and $\check{\mathbf{B}}_{y_{id}} = (\mathbf{Z}'_{s_{id}}\mathbf{L}_{s_{id}}\mathbf{Z}_{s_{id}})^{-1}\mathbf{Z}'_{s_{id}}\mathbf{L}_{s_{id}}\mathbf{Y}_{s_{id}}$ and $\check{\mathbf{B}}_d = \mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_{2d}}\mathbf{Z}_{s_{2d}}(\mathbf{Z}'_{s_{1d}}\mathbf{L}_{s_{1d}}\mathbf{Z}_{s_{1d}} + \mathbf{Z}'_{s_{2d}}\mathbf{L}_{s_{2d}}\mathbf{Z}_{s_{2d}})^{-1}$ with $\mathbf{L}_{s_{id}} = \mathbf{\Lambda}_{s_i}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{s_{id}}})$. The expressions for the approximate design variance of $\check{Y}^{\mathrm{CR}}_{1d}$ and $\check{\mathbf{Z}}^{\mathrm{CR}}_{1d}$ are exactly as the expressions (18) and (19) respectively, except that they now involve the estimators $\check{Y}^{\mathrm{R}}_{1d}$, $\check{\mathbf{Z}}^{\mathrm{R}}_{1d}$ and $\check{\mathbf{Z}}^{\mathrm{R}}_{2d}$ (in place of $\hat{Y}^{\mathrm{R}}_{1d}$, $\hat{\mathbf{Z}}^{\mathrm{R}}_{1d}$ and $\hat{\mathbf{Z}}^{\mathrm{R}}_{2d}$), and $\mathbf{B}_{y_1 d}$ and $\mathbf{B}_d$ are the population counterparts of $\check{\mathbf{B}}_{y_1 d}$ and $\check{\mathbf{B}}_d$ respectively.

Results that are identical with those of proposition 1 hold for estimators (23) and (24); the proof is similar to that of proposition 1. Estimators (23) and (24) are expected to be highly efficient because the regression on $\mathbf{x}_1$ and $\mathbf{x}_2$ is at the domain level. These results can be generalized to any number of domains in an obvious way, following the remarks at the end of Sections 2.2.2 and 3.1. 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*

For the same reasons as given in Section 2.2.2, it may be more sensible to use as auxiliary information at the domain level only the domain size, using the set-up

$$\mathcal{X}_s = \begin{pmatrix} \boldsymbol{\chi}_{s_1} & \mathbf{0} & \mathbf{Z}_{s_{1d}} \\ \mathbf{0} & \boldsymbol{\chi}_{s_2} & -\mathbf{Z}_{s_{2d}} \end{pmatrix}, \qquad \mathbf{t} = \begin{pmatrix} \mathbf{t}_{\chi_1} \\ \mathbf{t}_{\chi_2} \\ \mathbf{0} \end{pmatrix}, \tag{25}$$

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

$$\begin{aligned} \tilde{Y}^{\mathrm{CR}}_{1d} &= \tilde{Y}^{\mathrm{R}}_{1d} + \tilde{\mathbf{B}}'_{y_{1d}}(\mathbf{I} - \tilde{\mathbf{B}}_d)(\tilde{\mathbf{Z}}^{\mathrm{R}}_{2d} - \tilde{\mathbf{Z}}^{\mathrm{R}}_{1d}), \\ \tilde{Y}^{\mathrm{CR}}_{2d} &= \tilde{Y}^{\mathrm{R}}_{2d} - \tilde{\mathbf{B}}'_{y_{2d}}\tilde{\mathbf{B}}_d(\tilde{\mathbf{Z}}^{\mathrm{R}}_{2d} - \tilde{\mathbf{Z}}^{\mathrm{R}}_{1d}), \end{aligned} \tag{26}$$

and

$$\tilde{\mathbf{Z}}^{\mathrm{CR}}_{1d} = \tilde{\mathbf{Z}}^{\mathrm{CR}}_{2d} = \tilde{\mathbf{B}}_d\tilde{\mathbf{Z}}^{\mathrm{R}}_{1d} + (\mathbf{I} - \tilde{\mathbf{B}}_d)\tilde{\mathbf{Z}}^{\mathrm{R}}_{2d}, \tag{27}$$

where $\tilde{Y}^{\mathrm{R}}_{id}$ and $\tilde{\mathbf{Z}}^{\mathrm{R}}_{id}$ are GREG domain estimators of the type shown in expression (10), and $\tilde{\mathbf{B}}_{y_{id}} = (\mathbf{Z}'_{s_{id}}\boldsymbol{\mathcal{L}}_{s_i}\mathbf{Z}_{s_{id}})^{-1}\mathbf{Z}'_{s_{id}}\boldsymbol{\mathcal{L}}_{s_i}\mathbf{Y}_{s_{id}}$ and $\tilde{\mathbf{B}}_d = \mathbf{Z}'_{s_{2d}}\boldsymbol{\mathcal{L}}_{s_2}\mathbf{Z}_{s_{2d}}(\mathbf{Z}'_{s_{1d}}\boldsymbol{\mathcal{L}}_{s_1}\mathbf{Z}_{s_{1d}} + \mathbf{Z}'_{s_{2d}}\boldsymbol{\mathcal{L}}_{s_2}\mathbf{Z}_{s_{2d}})^{-1}$ with $\boldsymbol{\mathcal{L}}_{s_i} = \mathbf{\Lambda}_{s_i}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\chi}_{s_i}})$. We ensure, of course, that $(\boldsymbol{\chi}_{s_i}\ \mathbf{Z}_{s_{id}})$ is of full rank $p_i + 1 + q$. The expressions for the approximate design variance of $\tilde{Y}^{\mathrm{CR}}_{1d}$ and $\tilde{\mathbf{Z}}^{\mathrm{CR}}_{1d}$ are exactly as the expressions (18) and (19) respectively, except that they now involve the estimators $\tilde{Y}^{\mathrm{R}}_{1d}$, $\tilde{\mathbf{Z}}^{\mathrm{R}}_{1d}$ and $\tilde{\mathbf{Z}}^{\mathrm{R}}_{2d}$ (in place of $\hat{Y}^{\mathrm{R}}_{1d}$, $\hat{\mathbf{Z}}^{\mathrm{R}}_{1d}$ and $\hat{\mathbf{Z}}^{\mathrm{R}}_{2d}$), and $\mathbf{B}_{y_{1d}}$ and $\mathbf{B}_d$ are the population counterparts of $\mathbf{B}_{y_{1d}}$ and $\mathbf{B}_d$ respectively.

Results that are identical to those of proposition 1 hold for estimators (26) and (27); the proof is similar to that of proposition 1. Here again, a generalization to more than one domain is straightforward, following relevant remarks in Sections 2.2.3 and 3.1.

3.4. Comparing the three composite regression domain estimators

Just as with the comparison of the single-sample domain estimators in theorem 1, the three composite estimators (17), (24) and (27) are compared with respect to their approximate variance under the conditions of the following theorem; the proof is given in Appendix A.

Theorem 2.

(a) Suppose that $\mathbf{1} = \mathbf{X}_i \mathbf{h}_i$, for constant p_i-vectors $\mathbf{h}_i$. For both surveys, assume simple random sampling without replacement with sampling fractions $f_i = n_i/N$ and specify the constants q_{ik} in $\mathbf{\Lambda}_{s_i}$ as $q_{ik} = n_i/(1 - f_i)$. Then the following inequalities hold.

$$\mathrm{AV}(\check{\mathbf{Z}}_{id}^{\mathrm{CR}}) \leqslant \mathrm{AV}(\tilde{\mathbf{Z}}_{id}^{\mathrm{CR}}) \leqslant \mathrm{AV}(\hat{\mathbf{Z}}_{id}^{\mathrm{CR}}). \tag{28}$$

(b) For each survey assume stratified simple random sampling without replacement and combined regression in which the auxiliary row vector corresponding to unit k for survey i is $\mathbf{x}'_{ik} = (\mathbf{x}'_{0ik}, d_{1k}, \ldots, d_{H_i k})$, where $\mathbf{x}'_{0i}$ denotes the row p_i-vector of the core auxiliary variables and d_{lk} is the indicator variable of the membership of population unit k to stratum $l = 1, \ldots, H_i$. Then, if the constants q_{ik} in $\mathbf{\Lambda}_{s_i}$ are specified as $q_{ik} = n_{il}(N_{il} - 1)/N_{il}^2(1 - f_{il})$ for all units of stratum l, the inequalities in expression (28) hold.

(c) For both surveys assume Poisson sampling and specify the constants q_{ik} in $\mathbf{\Lambda}_{s_i}$ as $q_{ik} = \pi_{ik}/(1 - \pi_{ik})$. Then the inequality $\mathrm{AV}(\check{\mathbf{Z}}_{id}^{\mathrm{CR}}) \leqslant \mathrm{AV}(\tilde{\mathbf{Z}}_{id}^{\mathrm{CR}})$ holds under the condition $\mathbf{1} = \mathbf{X}_i \mathbf{h}_i$; the other inequalities in expression (28) hold without this condition.

Expression (28) shows that the variances of the three composite GREG estimators of the domain total $\mathbf{t}_{\mathbf{z}_d}$ preserve the ordering of the variances of their single-sample GREG counterparts (see expression (12)). Thus, the efficiency of the GREG composite estimators of $\mathbf{t}_{\mathbf{z}_d}$ depends on the amount of the incorporated domain-specific auxiliary information. Again, this conclusion is valid under the conditions of theorem 2, but it is expected to hold under general designs. An empirical study involving two samples with complex designs is presented next.

4. An empirical study

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 by rotation group into two subsamples, s_1 and s_2, of three rotations each. The six rotations that comprise the Labour Force Survey 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, we chose as 'common' variables to the two subsamples the Labour Force status categories 'employed' and 'unemployed'; the first of these two population characteristics is much more prevalent than the second. This setting simulates the case (b)(i) in Section 1. Regression was carried out separately for two Canadian provinces, namely Ontario and Saskatchewan, with the same calibration scheme for s_1 and s_2 involving population totals for seven age groups cross-classified with gender. The extended regression procedure for composite estimation incorporated the two common variables simultaneously. Four small geographic areas in each of the two provinces were used as study domains. The domain size, which is defined as the non-institutional population of people of age 15 years and over, ranges from 100 207 to 326 853

Table 1. Relative efficiencies of the domain estimators $\hat{Z}_d^{\mathrm{R}}$, $\tilde{Z}_d^{\mathrm{R}}$, $\hat{Z}_d^{\mathrm{CR}}$ and $\tilde{Z}_d^{\mathrm{CR}}$

Sample	*Area*	$\hat{V}(\hat{Z}_d^{\mathrm{R}})/\hat{V}(\tilde{Z}_d^{\mathrm{R}})$		$\hat{V}(\hat{Z}_d^{\mathrm{R}})/\hat{V}(\hat{Z}_d^{\mathrm{CR}})$		$\hat{V}(\tilde{Z}_d^{\mathrm{R}})/\hat{V}(\hat{Z}_d^{\mathrm{CR}})$		$\hat{V}(\tilde{Z}_d^{\mathrm{R}})/\hat{V}(\tilde{Z}_d^{\mathrm{CR}})$		$\hat{V}(\hat{Z}_d^{\mathrm{CR}})/\hat{V}(\tilde{Z}_d^{\mathrm{CR}})$		$\hat{V}(\hat{Z}_d^{\mathrm{R}})/\hat{V}(\tilde{Z}_d^{\mathrm{CR}})$	
		Results for Ontario											
		empl	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*
s_1	1	4.04	1.41	1.45	1.53	0.36	1.08	1.82	1.63	5.08	1.51	7.35	2.30
	2	3.88	1.01	1.83	1.37	0.47	1.35	1.44	1.65	3.06	1.22	5.59	1.66
	3	6.29	1.20	2.55	1.28	0.40	1.06	1.89	1.34	4.66	1.26	11.85	1.62
	4	11.70	1.19	2.63	1.84	0.23	1.54	2.64	2.19	11.74	1.42	30.91	2.61
s_2	1	5.91	1.59	2.58	2.28	0.44	1.43	2.22	2.16	5.08	1.51	13.11	3.44
	2	2.95	1.49	2.25	2.80	0.76	1.87	2.33	2.27	3.06	1.22	6.88	3.40
	3	3.84	1.32	1.64	3.45	0.43	2.61	1.99	3.30	4.66	1.26	7.64	4.35
	4	13.83	1.45	1.46	1.70	0.11	1.17	1.24	1.66	11.74	1.42	17.15	2.41
		prate	*urate*	*prate*	*urate*	*prate*	*urate*	*prate*	*urate*	*prate*	*urate*	*prate*	*urate*
s_1	1	0.93	0.98	1.26	1.60	1.35	1.62	1.97	1.59	1.46	0.98	1.84	1.57
	2	0.94	1.01	1.28	1.85	1.36	1.83	1.39	1.80	1.02	0.98	1.31	1.82
	3	0.92	1.00	1.22	1.33	1.33	1.33	1.76	1.30	1.32	0.98	1.62	1.30
	4	0.83	0.98	1.35	2.20	1.62	2.25	2.49	2.25	1.53	1.00	2.07	2.20
s_2	1	0.87	0.99	1.43	2.28	1.66	2.31	1.97	2.26	1.19	0.98	1.71	2.24
	2	0.98	0.99	1.20	2.07	1.23	2.10	2.44	2.06	1.99	0.98	2.39	2.04
	3	0.96	1.01	1.26	3.45	1.31	3.41	2.20	3.34	1.68	0.98	2.11	3.38
	4	0.79	1.00	1.31	1.68	1.66	1.68	1.33	1.68	0.80	1.00	1.05	1.68
		Results for Saskatchewan											
		empl	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*
s_1	1	3.81	0.87	1.75	1.08	0.46	1.24	2.18	1.22	4.75	0.99	8.31	1.07
	2	3.33	0.71	1.07	2.32	0.32	3.25	2.33	2.46	7.24	0.76	7.76	1.75
	3	5.63	1.19	1.09	3.01	0.19	2.53	1.79	2.72	9.24	1.07	10.09	3.24
	4	4.24	1.32	1.96	1.70	0.46	1.29	1.62	1.51	3.51	1.17	6.86	1.99

(*continued*)

Table 1. (*continued*)

Sample	*Area*	$\hat{V}(\hat{Z}_d^{\mathrm{R}})/\hat{V}(\tilde{Z}_d^{\mathrm{R}})$		$\hat{V}(\hat{Z}_d^{\mathrm{R}})/\hat{V}(\hat{Z}_d^{\mathrm{CR}})$		$\hat{V}(\tilde{Z}_d^{\mathrm{R}})/\hat{V}(\hat{Z}_d^{\mathrm{CR}})$		$\hat{V}(\tilde{Z}_d^{\mathrm{R}})/\hat{V}(\tilde{Z}_d^{\mathrm{CR}})$		$\hat{V}(\hat{Z}_d^{\mathrm{CR}})/\hat{V}(\tilde{Z}_d^{\mathrm{CR}})$		$\hat{V}(\hat{Z}_d^{\mathrm{R}})/\hat{V}(\tilde{Z}_d^{\mathrm{CR}})$	
		Results for Saskatchewan											
		empl	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*	*empl*	*unempl*
s_2	1	6.04	1.18	2.31	3.19	0.38	2.72	1.81	2.69	4.75	0.99	10.95	3.17
	2	14.91	0.69	3.18	1.49	0.21	2.14	1.54	1.62	7.24	0.76	23.02	1.12
	3	12.72	0.96	3.08	0.94	0.24	0.98	2.24	1.05	9.24	1.07	28.47	1.01
	4	3.03	0.94	1.90	2.33	0.63	2.49	2.20	2.92	3.51	1.17	6.66	2.73
		prate	*urate*	*prate*	*urate*	*prate*	*urate*	*prate*	*urate*	*prate*	*urate*	*prate*	*urate*
s_1	1	0.95	1.00	1.42	1.27	1.49	1.27	2.13	1.27	1.43	1.00	2.03	1.27
	2	0.98	1.03	1.15	2.36	1.17	2.29	2.55	2.29	2.18	1.00	2.51	2.36
	3	0.96	0.98	1.00	2.73	1.04	2.78	1.60	2.78	1.54	1.00	1.54	2.73
	4	0.96	0.96	1.05	1.42	1.09	1.48	1.51	1.48	1.38	1.00	1.45	1.42
s_2	1	0.99	1.01	1.23	2.60	1.24	2.57	1.95	2.57	1.57	1.00	1.94	2.60
	2	0.91	0.95	1.12	1.64	1.23	1.73	1.32	1.73	1.08	1.00	1.21	1.64
	3	0.99	0.97	0.99	1.03	1.00	1.06	2.30	1.06	2.30	1.00	2.27	1.03
	4	0.99	0.98	1.51	2.82	1.54	2.86	2.50	2.86	1.63	1.00	2.47	2.82

in the four Ontario areas, and from 83066 to 184407 in the four Saskatchewan areas. The one-sample domain estimators $\hat{Z}_d^{\mathrm{R}}$ and $\tilde{Z}_d^{\mathrm{R}}$ and their composite counterparts $\hat{Z}_d^{\mathrm{CR}}$ and $\tilde{Z}_d^{\mathrm{CR}}$ (for scalar characteristics) were compared with respect to their estimated (by the jackknife method) variances; the subscript i in the notation of these estimators is omitted here for simplicity of exposition. The main results are displayed in Table 1, where $\hat{V}$ denotes estimated variance.

Starting with the one-sample estimators $\hat{Z}_d^{\mathrm{R}}$ and $\tilde{Z}_d^{\mathrm{R}}$, the third column in Table 1 shows huge gains in efficiency when using the estimator $\tilde{Z}_d^{\mathrm{R}}$ instead of $\hat{Z}_d^{\mathrm{R}}$ for the number of employed people (empl); this corresponds to the middle part of expression (12), but under the complex design of the Labour Force Survey. The high efficiency of $\tilde{Z}_d^{\mathrm{R}}$ for empl is due to the strong correlation of the binary characteristic employed with the auxiliary vector $\mathbf{x}$. This strong correlation is implied by the high prevalence of employed and the fact that this $\mathbf{x}$ satisfies $1 = \mathbf{x}_{\mathbf{k}}'\mathbf{h}$ for $k \in s$. The differences in gain in efficiency between areas are explained by the varying degree of prevalence of employed (e.g. highest efficiency for highest prevalence in area 4 in Ontario), though they may be partly due to the varying correlation of employed with the age and sex regressors. The fourth column shows much smaller, yet substantial, gains for the number of unemployed people (unempl) for most areas in Ontario, whereas a loss of efficiency is observed in half of the areas in Saskatchewan; this supports the argument (see Section 2.2.1) that the effectiveness of regression in variance reduction weakens when we estimate low prevalence (rare) characteristics. Interestingly, $\tilde{Z}_d^{\mathrm{R}}$ is less efficient than $\hat{Z}_d^{\mathrm{R}}$ for the unemployment rate (urate), which is defined as unempl/(unempl + empl), and even less efficient for the participation rate (prate), which is defined as (unempl + empl)/domain size; in most cases, though, the loss of efficiency is slight. For prate this can be explained by the greater stability of $\hat{Z}_d^{\mathrm{R}}$, which is a ratio of two estimators, whereas only the numerator of $\tilde{Z}_d^{\mathrm{R}}$ is variable, the denominator being the controlled domain size; this is akin to the stability of the Hajek estimator of a mean. The explanation for urate is along these lines.

In contrast, the fifth and sixth columns show that the composite estimator $\hat{Z}_d^{\mathrm{CR}}$ is much more efficient than the one-sample estimators $\hat{Z}_d^{\mathrm{R}}$, based on either s_1 or s_2, for both empl and unempl; this corresponds to expression (20), but under the complex design of the Labour Force Survey. The only exception is a slight loss in area 3, sample s_2, in Saskatchewan. $\hat{Z}_d^{\mathrm{CR}}$ is also more efficient than $\hat{Z}_d^{\mathrm{R}}$ for prate and even more for urate.

The third and fourth, and fifth and sixth columns show that controlling the domain size, as in $\tilde{Z}_d^{\mathrm{R}}$, is more efficient (relative to $\hat{Z}_d^{\mathrm{R}}$) than using information from another sample, as in $\hat{Z}_d^{\mathrm{CR}}$, for empl, whereas the reverse is true for unempl. This becomes very clear in the seventh and eighth columns, which show the direct comparison of $\tilde{Z}_d^{\mathrm{R}}$ with $\hat{Z}_d^{\mathrm{CR}}$—controlling the domain size is far more efficient than combining information from s_1 and s_2 for empl, whereas the reverse is true for unempl. Importantly, combining information is more efficient than controlling the domain size for prate and more so for urate; contrast this with the third and fourth columns. The ninth and 10th columns show that combining information in addition to controlling the domain size improves substantially the efficiency of the domain estimators for both empl and unempl. This holds also for prate and urate. This comparison is analogous to that shown in the fifth and sixth columns, but more revealing, in that it demonstrates that combining information is beneficial even when auxiliary information at the domain level has been used.

Regarding the composite estimators $\tilde{Z}_d^{\mathrm{CR}}$ and $\hat{Z}_d^{\mathrm{CR}}$, the 11th and 12th columns show that impressive gains in efficiency result from using domain size controls in addition to combining information from the two samples. This is an empirical illustration, but for the complex survey situation studied, of the result of theorem 2. The gains are much larger for empl than for unempl because controlling the domain size is less effective for the latter low prevalence characteristic. In Saskatchewan we observe loss of efficiency for unempl, but in fewer areas and of lesser

degree than in the one-sample counterparts of $\tilde{Z}_d^{\mathrm{CR}}$ and $\hat{Z}_d^{\mathrm{CR}}$ that are compared in the third and fourth columns. Interestingly, controlling the domain size in the presence of information from the other sample turns the slight loss of efficiency for prate that is caused by using the estimator $\tilde{Z}_d^{\mathrm{R}}$ instead of $\hat{Z}_d^{\mathrm{R}}$ into considerable gain; for urate we observe a slight loss in Ontario, which is comparable with that shown in the third and fourth columns, whereas no effect is observed in Saskatchewan.

The combined effect of using information on empl and unempl from the two samples and controlling the domain size (indicated by comparing $\hat{Z}_d^{\mathrm{R}}$ with $\tilde{Z}_d^{\mathrm{CR}}$) is shown in the 13th and 14th columns; the separate effects are shown in the fifth and sixth, and 11th and 12th columns. Very large gains are realized for both empl and unempl, much more so for empl, and also for prate and urate. Contrasting the third and fourth columns and 13th and 14th makes evident the highly beneficial effect of borrowing strength from another survey even when domain size information has been used in the domain estimation.

As with empl and unempl, for other categorical characteristics within the geographical areas the gain in efficiency (which is not tabulated here) resulting from controlling the area size depends on their prevalence. Although the gain is huge for high prevalence characteristics, it diminishes with decreasing prevalence, and for characteristics of very low prevalence we observed a substantial loss of efficiency in some areas (mainly in Saskatchewan). Similarly, borrowing strength from the other sample is more effective for the non-common characteristics of higher prevalence, especially for those correlated with empl. In contrast, for characteristics that are weakly correlated with empl, including characteristics of lower prevalence than empl, the combination of information on empl from the two samples results in a moderate gain, or even loss, of efficiency. It is worth noting that, with few exceptions involving low prevalence characteristics, controlling the area size was more efficient than borrowing strength from the other survey. This is because in the former procedure we use as control total the non-random domain size, whereas in the latter we use estimates for the common variables from the other survey. Furthermore, for non-common variables the comparison of the composite estimators $\hat{Y}_d^{\mathrm{CR}}$ and $\tilde{Y}_d^{\mathrm{CR}}$, for which there is no analytic result that is analogous to inequality (28), showed that $\tilde{Y}_d^{\mathrm{CR}}$ is substantially more efficient than $\hat{Y}_d^{\mathrm{CR}}$—more so for high prevalence characteristics.

Similar were the findings, which are not displayed here, when s_1 and s_2 were made up of two and four rotations respectively. As expected, the distinctive feature in this situation was that the gains in efficiency for $\hat{Z}_d^{\mathrm{CR}}$ relative to $\hat{Z}_{1d}^{\mathrm{R}}$ were larger than relative to $\hat{Z}_{2d}^{\mathrm{R}}$, as more information flowed from s_2 to s_1 than from s_1 to s_2. For the same reason, the gains in efficiency for $\tilde{Z}_d^{\mathrm{CR}}$ relative to $\tilde{Z}_{1d}^{\mathrm{R}}$ were larger than relative to $\tilde{Z}_{2d}^{\mathrm{R}}$.

5. Summary and discussion

Analytical results for certain sampling designs show that using auxiliary information at the domain level, if it exists and when suitable, improves the precision of domain estimates. Extending the regression procedure so as to combine comparable information from two surveys at the domain level improves the precision even further: substantially for common survey variables but less so for non-common variables. An empirical study of the effect of combining at the domain level two samples with complex designs has provided a quantification of the resulting gains in efficiency for estimated totals of common and non-common 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 characteristic 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 are not based on repeated samples and may therefore be regarded as strongly suggestive but not conclusive.

It was shown in Section 3 that the GREG procedure proposed 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 non-common, 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 (14), or (22) or (25) with the submatrix $(\mathbf{Z}_{s_1}, -\mathbf{Z}_{s_2})'$. This will be redundant if the domains that are included in the procedure are exhaustive.

The number of common variables for which we seek domain estimates may be so large 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 quite customary in small area 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 control totals are ineffective for domain estimation, it is prudent to use only domain level control totals.

In the special situation when all variables are common between the two surveys (as when one of the surveys is supplementary to 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 expressions (15), (23) or (26), which can be combined as already described in Section 3.1 for the estimators in expression (15). In the complex survey setting of the empirical study, the estimated variances of both $\hat{Y}_d^{\mathrm{CR}}$ and $\phi\hat{Y}_{1d}^{\mathrm{R}} + (1-\phi)\hat{Y}_{2d}^{\mathrm{R}}$ were computed for several 'non-common' variables. The variance of $\hat{Y}_d^{\mathrm{CR}}$ was smaller for most of those variables, substantially so for some.

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 is too small. However, even domain sizes may not be available for some small domains, and most likely not for non-geographic domains. Nevertheless, 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 (see the third and fourth and 11th and 12th columns of Table 1) owing to the 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. Of course, the successful implementation of the method proposed rests on the assumed comparability of totals for variables that are common to the combined surveys. To the extent that the comparability of these totals is only approximate, sufficient harmonization being unworkable, the case for composite estimation in reducing mean-squared error is less compelling, but it may still be beneficial.

If the size of the observed difference between estimates of comparable items cannot be explained by sampling variability, the common questionnaire items between the two surveys

may be regarded as estimating somewhat different population quantities owing to reasons that were mentioned in Section 1. In such cases a sensible approach is as follows. Assume that the sample of the primary survey, for which small domain estimates are sought, is s_1, whereas the sample from which strength is to be borrowed is s_2. Then, for the 'common' vector $\mathbf{z}$, we uphold the design-based property $E(\hat{\mathbf{Z}}_{1d}) = \mathbf{t}_{\mathbf{z}_d}$ of the HT estimator, but postulate that $E(\hat{\mathbf{Z}}_{2d}) = \mathbf{t}_{\mathbf{z}_d} + \mathbf{b}$, where $\mathbf{b}$ stands for the 'bias' in estimating $\mathbf{t}_{\mathbf{z}_d}$ using s_2. For large samples, the composite GREG estimator $\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}$ given by expression (17) has bias $(\mathbf{I} - \mathbf{B}_d)\mathbf{b}$ and approximate mean-squared error $\mathrm{AMSE}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) + ((\mathbf{I} - \mathbf{B}_d)\mathbf{b})((\mathbf{I} - \mathbf{B}_d)\mathbf{b})'$, with $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}})$ given by equation (19). Under the conditions of proposition 1, when the exact form of $\mathbf{B}_d$ can be worked out, the loss of efficiency when ignoring the bias is

$$\{\mathrm{AMSE}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) - \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}})\}\,\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}})^{-1} = \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})C\mathbf{b}\mathbf{b}'C\,\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})C\,\mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1},$$

where $C = \{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}$. For a scalar z, this loss has the simple form $\mathrm{AV}(\hat{Z}_{1d}^{\mathrm{R}})b^2 \times [\mathrm{AV}(\hat{Z}_{2d}^{\mathrm{R}})\mathrm{AV}(\hat{Z}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{Z}_{2d}^{\mathrm{R}})\}]^{-1}$. Note that, the smaller $\mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})$ is, the larger the loss of efficiency. A simple calculation, under the conditions of proposition 1, shows that, if $\mathbf{b} > \{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{1/2}$, then $\mathrm{AMSE}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) > \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})$, i.e. ignoring a bias of this size results in a composite GREG estimator that is less efficient than its single-sample counterpart. A modification of the regression set-up (14) that corrects for the bias $\mathbf{b}$ involves the adjusted vector of control totals $\mathbf{t} = (\mathbf{t}_{\mathbf{x}_1}', \mathbf{t}_{\mathbf{x}_2}', -\mathbf{b}')$. This results in the approximately unbiased composite GREG estimator $\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}} = \hat{\mathbf{B}}_d\hat{\mathbf{Z}}_{1d}^{\mathrm{R}} + (\mathbf{I} - \hat{\mathbf{B}}_d)(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}} - \mathbf{b})$, with $\hat{\mathbf{B}}_d$ as in equation (17) and having the approximate variance $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}})$ given by equation (19). In practice, an estimate of the unknown bias that can be used in the vector $\mathbf{t}$ is $\hat{\mathbf{b}} = \hat{\mathbf{Z}}_{2d} - \hat{\mathbf{Z}}_{1d}$. If some of the components of the vector $\hat{\mathbf{b}}$ are tested to be statistically significant, they can be used as non-zero entries of $\hat{\mathbf{b}}$ in $\mathbf{t}$ and be treated as constants like the other components of $\mathbf{t}$. An alternative composite GREG estimator involves the weighting coefficient $\hat{\mathbf{D}}_d = \{\widehat{\mathrm{AV}}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}) + \hat{\mathbf{b}}\hat{\mathbf{b}}'\}\{\widehat{\mathrm{AV}}(\hat{Z}_{1d}^{\mathrm{R}}) + \widehat{\mathrm{AV}}(\hat{Z}_{2d}^{\mathrm{R}}) + \hat{\mathbf{b}}\hat{\mathbf{b}}'\}^{-1}$. Given in terms of estimated quantities, $\hat{\mathbf{D}}_d$ minimizes the estimated approximate mean-squared error of $\hat{\mathbf{D}}_d\hat{\mathbf{Z}}_{1d}^{\mathrm{R}} + (\mathbf{I} - \hat{\mathbf{D}}_d)\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}$. Notably, this estimator is not a calibration estimator.

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

Acknowledgements

The author thanks the Joint Editor, Associate Editor and three referees for their comments, which have substantially improved the paper. Most of this research was carried out at Statistics Canada.

Appendix A: Proofs

A.1. Proof of theorem 1

(a) For simple random sampling with sampling fraction $f = n/N$, it can be easily shown that $\mathbf{\Lambda}^0 = \lambda^0(\mathbf{I} - \mathbf{P_1})$, where $\lambda^0 = N^2(1 - f)/n(N - 1)$ and $\mathbf{P_1} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$. Then, it follows from $\mathbf{1} = \mathbf{X}\mathbf{h}$ that $\mathbf{P_X}\mathbf{P_1} = \mathbf{P_1}$ and $\mathbf{P_X} - \mathbf{P_1} = (\mathbf{P_X} - \mathbf{P_1})^2$, so equation (5) can be written as

$$\begin{aligned}\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}}) &= \lambda^0\mathbf{Y}_d'(\mathbf{I} - \mathbf{P_1})\mathbf{Y}_d - \lambda^0\mathbf{Y}_d'(\mathbf{P_X} - \mathbf{P_1})\mathbf{Y}_d \\ &= V(\hat{\mathbf{Y}}_d) - \lambda^0\mathbf{Y}_d'(\mathbf{P_X} - \mathbf{P_1})^2\mathbf{Y}_d,\end{aligned}$$

which implies that $\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}}) \leqslant V(\hat{\mathbf{Y}}_d)$.

Next, from $(\mathbf{I}-\mathbf{P_X})\boldsymbol{\Lambda}^0(\mathbf{I}-\mathbf{P_X})=\lambda^0(\mathbf{I}-\mathbf{P_X})$ we obtain $\mathbf{Y}_d'\mathbf{L}\mathbf{1}_d(\mathbf{1}_d'\mathbf{L}\mathbf{1}_d)^{-1}=\mathrm{AC}(\hat{\mathbf{Y}}_d^{\mathrm{R}},\hat{N}_d^{\mathrm{R}})\,\mathrm{AV}(\hat{N}_d^{\mathrm{R}})^{-1}$, and thus equation (11) becomes

$$\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})=\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})-\mathrm{AC}(\hat{\mathbf{Y}}_d^{\mathrm{R}},\hat{N}_d^{\mathrm{R}})\,\mathrm{AV}(\hat{N}_d^{\mathrm{R}})^{-1}(\mathrm{AC}(\hat{\mathbf{Y}}_d^{\mathrm{R}},\hat{N}_d^{\mathrm{R}}))'.$$

This shows that $\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})\leqslant\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})$.

Finally, in view of $\boldsymbol{\Lambda}^0=\lambda^0(\mathbf{I}-\mathbf{P_1})$ and $(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})^2=(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})$, $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})$ in equation (8) can be expanded as

$$\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})=\lambda^0\mathbf{Y}_d'\{(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})-(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})\mathbf{P_1}(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})\}\mathbf{Y}_d.$$

Noting that $\mathbf{Y}_d'\mathbf{1}=\mathbf{Y}_d'\mathbf{1}_d$, $\mathbf{X}_d'\mathbf{1}=\mathbf{X}_d'\mathbf{1}_d$ and that $\mathbf{1}=\mathbf{X}\mathbf{h}$ implies that $\mathbf{1}_d=\mathbf{X}_d\mathbf{h}$, it is trivial to show that $\mathbf{Y}_d'(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})\mathbf{P_1}(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})\mathbf{Y}_d=\mathbf{0}$, from which it follows that $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})=\lambda^0\mathbf{Y}_d'(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})\mathbf{Y}_d$. Moreover, expression (11) reduces to

$$\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})=\lambda^0\mathbf{Y}_d'[\mathbf{I}-\mathbf{P_X}-(\mathbf{I}-\mathbf{P_X})\mathbf{1}_d\{\mathbf{1}_d'(\mathbf{I}-\mathbf{P_X})\mathbf{1}_d\}^{-1}\mathbf{1}_d'(\mathbf{I}-\mathbf{P_X})]\mathbf{Y}_d=\lambda^0\mathbf{Y}_d'(\mathbf{I}-\mathbf{P}_{\boldsymbol{\mathcal{X}}})\mathbf{Y}_d,$$

in obvious notation for $\mathbf{P}_{\boldsymbol{\mathcal{X}}}$. Then $\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})-\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})=\lambda^0\mathbf{Y}_d'(\mathbf{P}_{\mathbf{X}_d}-\mathbf{P}_{\boldsymbol{\mathcal{X}}})\mathbf{Y}_d$, and to show that $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})\leqslant\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})$ it suffices to show that $\mathbf{Y}_d'(\mathbf{P}_{\mathbf{X}_d}-\mathbf{P}_{\boldsymbol{\mathcal{X}}})\mathbf{Y}_d=\mathbf{Y}_d'(\mathbf{P}_{\mathbf{X}_d}-\mathbf{P}_{\boldsymbol{\mathcal{X}}})^2\mathbf{Y}_d$. Now $(\mathbf{P}_{\mathbf{X}_d}-\mathbf{P}_{\boldsymbol{\mathcal{X}}})^2=\mathbf{P}_{\mathbf{X}_d}-\mathbf{P}_{\mathbf{X}_d}\mathbf{P}_{\boldsymbol{\mathcal{X}}}-\mathbf{P}_{\boldsymbol{\mathcal{X}}}\mathbf{P}_{\mathbf{X}_d}+\mathbf{P}_{\boldsymbol{\mathcal{X}}}$. Making use of $\mathbf{1}_d=\mathbf{X}_d\mathbf{h}$, $\mathbf{X}_d'\mathbf{X}=\mathbf{X}_d'\mathbf{X}_d$ and $\mathbf{Y}_d'\mathbf{X}=\mathbf{Y}_d'\mathbf{X}_d$ we obtain after some algebra that $\mathbf{Y}_d'\mathbf{P}_{\mathbf{X}_d}\mathbf{P}_{\boldsymbol{\mathcal{X}}}\mathbf{Y}_d=\mathbf{Y}_d'\mathbf{P}_{\boldsymbol{\mathcal{X}}}\mathbf{Y}_d$. Similarly, $\mathbf{Y}_d'\mathbf{P}_{\boldsymbol{\mathcal{X}}}\mathbf{P}_{\mathbf{X}_d}\mathbf{Y}_d=\mathbf{Y}_d'\mathbf{P}_{\boldsymbol{\mathcal{X}}}\mathbf{Y}_d$. Therefore, $\mathbf{Y}_d'(\mathbf{P}_{\mathbf{X}_d}-\mathbf{P}_{\boldsymbol{\mathcal{X}}})^2\mathbf{Y}_d=\mathbf{Y}_d'(\mathbf{P}_{\mathbf{X}_d}-\mathbf{P}_{\boldsymbol{\mathcal{X}}})\mathbf{Y}_d\geqslant\mathbf{0}$.

(b) Here, $\boldsymbol{\Lambda}^0=\mathrm{diag}(\boldsymbol{\Lambda}_l^0)=\mathrm{diag}(\lambda_l^0\mathbf{I})\{\mathbf{I}-\mathrm{diag}(\mathbf{P}_{\mathbf{1}_l})\}$, where $\lambda_l^0=N_l^2(1-f_l)/n_l(N_l-1)$, and $\mathbf{X}=(\mathbf{X}_0\ \mathbf{D})$, where $\mathbf{D}$ denotes the $N\times H$ matrix whose lth column is the indicator vector $\mathbf{1}_l$ of unit membership to stratum l. Observe that $\mathrm{diag}(\lambda_l^0\mathbf{I})=\mathbf{Q}\,(=\mathrm{diag}(q_l^{-1}\mathbf{I}))$, and write $\mathbf{P_D}=\mathbf{D}(\mathbf{D}'\mathbf{Q}\mathbf{D})^{-1}\mathbf{D}'\mathbf{Q}$. It can be easily checked that $\mathbf{P_D}=\mathrm{diag}(\mathbf{P}_{\mathbf{1}_l})$. Then, with $\mathbf{P_X}=\mathbf{X}(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}$,

$$\begin{aligned}\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})&=\mathbf{Y}_d'(\mathbf{I}-\mathbf{P}_{\mathbf{X}}')\boldsymbol{\Lambda}^0(\mathbf{I}-\mathbf{P_X})\mathbf{Y}_d\\&=\mathbf{Y}_d'(\mathbf{I}-\mathbf{P}_{\mathbf{X}}')\mathbf{Q}(\mathbf{I}-\mathbf{P_D})(\mathbf{I}-\mathbf{P_X})\mathbf{Y}_d.\end{aligned}$$

Using algebra of partitioned matrices, it can be shown that

$$\mathbf{P_X}=(\mathbf{I}-\mathbf{P_D})\mathbf{X}_0\{\mathbf{X}_0'\mathbf{Q}(\mathbf{I}-\mathbf{P_D})\mathbf{X}_0\}^{-1}\mathbf{X}_0'\mathbf{Q}+(\mathbf{I}-\mathbf{P}_{\mathbf{X}_0})\mathbf{D}\{\mathbf{D}'\mathbf{Q}(\mathbf{I}-\mathbf{P}_{\mathbf{X}_0})\mathbf{D}\}^{-1}\mathbf{D}'\mathbf{Q},$$

where $\mathbf{P}_{\mathbf{X}_0}=\mathbf{X}_0(\mathbf{X}_0'\mathbf{Q}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Q}$. Noting that $\mathbf{P_D}\mathbf{P_D}=\mathbf{P_D}$, it can be easily verified that $\mathbf{P_D}\mathbf{P_X}=\mathbf{P_D}$ and thus $\mathbf{P_D}(\mathbf{I}-\mathbf{P_X})=\mathbf{0}$. Also easy to show is that $(\mathbf{I}-\mathbf{P}_{\mathbf{X}}')\mathbf{Q}(\mathbf{I}-\mathbf{P_X})=\mathbf{Q}(\mathbf{I}-\mathbf{P_X})$. It follows then that

$$\begin{aligned}\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})&=\mathbf{Y}_d'(\mathbf{I}-\mathbf{P}_{\mathbf{X}}')\mathbf{Q}(\mathbf{I}-\mathbf{P_X})\mathbf{Y}_d\\&=\mathbf{Y}_d'\mathbf{Q}(\mathbf{I}-\mathbf{P_X})\mathbf{Y}_d.\end{aligned}$$

Using $\mathbf{P_D}\mathbf{P_X}=\mathbf{P_D}$ and noting that $\mathbf{Q}\mathbf{P_X}=\mathbf{P}_{\mathbf{X}}'\mathbf{Q}$ and $\mathbf{Q}\mathbf{P_D}=\mathbf{P}_{\mathbf{D}}'\mathbf{Q}$ we can write

$$\begin{aligned}\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})&=\mathbf{Y}_d'\mathbf{Q}(\mathbf{I}-\mathbf{P_D})\mathbf{Y}_d-\mathbf{Y}_d'\mathbf{Q}(\mathbf{P_X}-\mathbf{P_D})\mathbf{Y}_d\\&=V(\hat{\mathbf{Y}}_d)-\mathbf{Y}_d'\mathbf{Q}(\mathbf{P_X}-\mathbf{P_D})^2\mathbf{Y}_d,\end{aligned}$$

which implies that $\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})\leqslant V(\hat{\mathbf{Y}}_d)$, since $\mathbf{Q}$ is a diagonal matrix with positive diagonal entries.

In view of $\mathbf{L}=\mathbf{Q}(\mathbf{I}-\mathbf{P_X})$, the proof of $\mathrm{AV}(\tilde{Y}_d^{\mathrm{R}})\leqslant\mathrm{AV}(\hat{\mathbf{Y}}_d^{\mathrm{R}})$ is as in (a).

Next, $\mathbf{X}_d=(\mathbf{X}_{0d}\ \ \mathbf{D}_d)$, and $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})$ in equation (8) can be written as $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})=\mathbf{Y}_d'(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d}')\mathbf{Q}(\mathbf{I}-\mathbf{P_D})(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})\mathbf{Y}_d$. Using the expanded form of $\mathbf{P}_{\mathbf{X}_d}$, as above, and noting that $\mathbf{D}'\mathbf{X}_{0d}=\mathbf{D}_d'\mathbf{X}_{0d}$ and $\mathbf{D}'\mathbf{Y}_d=\mathbf{D}_d'\mathbf{Y}_d$, it can be easily shown that $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})=\mathbf{Y}_d'\mathbf{Q}(\mathbf{I}-\mathbf{P}_{\mathbf{X}_d})\mathbf{Y}_d$. Now, $\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})$ can be written as $\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})=\mathbf{Y}_d'\mathbf{Q}(\mathbf{I}-\mathbf{P}_{\boldsymbol{\mathcal{X}}})\mathbf{Y}_d$, where

$$\mathbf{P}_{\boldsymbol{\mathcal{X}}}=\mathbf{P_X}+(\mathbf{I}-\mathbf{P_X})\mathbf{1}_d\{\mathbf{1}_d'\mathbf{Q}(\mathbf{I}-\mathbf{P_X})\mathbf{1}_d\}^{-1}\mathbf{1}_d'\mathbf{Q}(\mathbf{I}-\mathbf{P_X}).$$

Noting that here, also, $\mathbf{1}_d=\mathbf{X}_d\mathbf{h}$, the proof of the inequality $\mathrm{AV}(\check{\mathbf{Y}}_d^{\mathrm{R}})\leqslant\mathrm{AV}(\tilde{\mathbf{Y}}_d^{\mathrm{R}})$ is as in (a).

(c) For Poisson sampling we have $\boldsymbol{\Lambda}^0=\mathrm{diag}(\pi_1^{-1}-1,\ldots,\pi_N^{-1}-1)$. Observing that here $\boldsymbol{\Lambda}^0=\mathbf{Q}$, the proof of the inequalities is a simpler version of the proof in (b)—simply ignore $\mathbf{D}$ in (b).

A.2. Proof of proposition 1

The proof will be given for $\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}$ and $\hat{Y}_{1d}^{\mathrm{R}}$; the proof for $\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}$ and $\hat{Y}_{2d}^{\mathrm{R}}$ is similar.

(a) Expression (19) can be rewritten as $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \mathbf{B}_d\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}\mathbf{B}_d' + (\mathbf{I} - \mathbf{B}_d)\,\mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}) - \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\mathbf{B}_d'$. Now, $\mathbf{L}_i = q_i^{-1}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})$, where $q_i = n_i/(1 - f_i)$ and, in view of the proof of theorem 1, $\mathrm{AV}(\hat{\mathbf{Z}}_{id}^{\mathrm{R}}) = \lambda_i^0\mathbf{Z}_d'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{Z}_d = \{N^2/(N-1)\}\mathbf{Z}_d'\mathbf{L}_i\mathbf{Z}_d$, from which it follows that $\mathbf{B}_d = \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}) \times \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}$. Then $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}})$ takes the form $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1} \times \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}) = \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) - \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}\,\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})$, which shows that $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) < \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})$.

Similarly, equation (18) can take the form

$$\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{CR}}) = \mathrm{AV}(\hat{Y}_{1d}^{\mathrm{R}}) - \mathrm{AC}(\hat{Y}_{1d}^{\mathrm{R}}, \hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}(\mathrm{AC}(\hat{Y}_{1d}^{\mathrm{R}}, \hat{\mathbf{Z}}_{1d}^{\mathrm{R}}))',$$

which shows that $\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{CR}}) < \mathrm{AV}(\hat{Y}_{1d}^{\mathrm{R}})$.

Now, $\mathbf{X}_1 = \mathbf{X}_2$ implies that $\mathbf{I} - \mathbf{P}_{\mathbf{X}_1} = \mathbf{I} - \mathbf{P}_{\mathbf{X}_2}$ and, thus, $\mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}) = (\lambda_2^0/\lambda_1^0)\,\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})$. Therefore, $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \{\lambda_2^0/(\lambda_1^0 + \lambda_2^0)\}\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})$, and $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}})\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} = \{\lambda_2^0/(\lambda_1^0 + \lambda_2^0)\}\mathbf{I} = \{n_1/(n_1 + n_2)\}\mathbf{I}$ under the condition $(1 - f_1)/(1 - f_2) \approx 1$.

Also,

$$\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{CR}}) = \mathrm{AV}(\hat{Y}_{1d}^{\mathrm{R}}) - \{\lambda_1^0/(\lambda_1^0 + \lambda_2^0)\}\,\mathrm{AC}(\hat{Y}_{1d}^{\mathrm{R}}, \hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\,\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1}(\mathrm{AC}(\hat{Y}_{1d}^{\mathrm{R}}, \hat{\mathbf{Z}}_{1d}^{\mathrm{R}}))'.$$

By the Cauchy–Schwarz inequality, $\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{CR}}) > \mathrm{AV}(\hat{Y}_{1d}^{\mathrm{R}}) - \{\lambda_1^0/(\lambda_1^0 + \lambda_2^0)\}\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{R}})$, and it follows that $\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{CR}})\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{R}})^{-1} > \lambda_2^0/(\lambda_1^0 + \lambda_2^0) = n_1/(n_1 + n_2)$, under the condition $(1 - f_1)/(1 - f_2) \approx 1$.

(b) Here, $\mathbf{L}_i = \mathbf{Q}_i(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})$, where $\mathbf{Q}_i = \mathrm{diag}(q_{il}^{-1}\mathbf{I})$ and $\mathbf{P}_{\mathbf{X}_i} = \mathbf{X}_i(\mathbf{X}_i'\mathbf{Q}_i\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Q}_i$. From theorem 1, $\mathrm{AV}(\hat{\mathbf{Z}}_{id}^{\mathrm{R}}) = \mathbf{Z}_d'\mathbf{Q}_i(\mathbf{I} - \mathbf{P}_{\mathbf{X}_i})\mathbf{Z}_d = \mathbf{Z}_d'\mathbf{L}_i\mathbf{Z}_d$, from which it follows that $\mathbf{B}_d = \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}$. The rest of the proof of expression (20) is as in (a). For the proof of expression (21), the assumptions imply that $\mathbf{Q}_i = q_i^{-1}\mathbf{I}$, where $q_i = f_i/(1 - f_i)$ and $f_i = n_i/N$. Then $\mathrm{AV}(\hat{\mathbf{Z}}_{id}^{\mathrm{R}}) = q_i^{-1}\mathbf{Z}_d'(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Z}_d$ and so $\mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}) = (q_1/q_2)\,\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})$ and $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \{q_1/(q_1 + q_2)\}\,\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) = \{n_1/(n_1 + n_2)\}\,\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})$, under the condition $(1 - f_1)/(1 - f_2) \approx 1$. It follows then that $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}})\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} = \{n_1/(n_1 + n_2)\}\mathbf{I}$. Similarly, $\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{CR}})\,\mathrm{AV}(\hat{Y}_{1d}^{\mathrm{R}})^{-1} > n_1/(n_1 + n_2)$.

(c) The proofs are exactly as in (b).

A.3. Proof of theorem 2

It suffices to give the proof for $\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}$, $\tilde{\mathbf{Z}}_{1d}^{\mathrm{CR}}$ and $\check{\mathbf{Z}}_{1d}^{\mathrm{CR}}$.

(a) It was shown in the proof of proposition 1 that $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}\mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})$. Analogous expressions can be derived for $\mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{CR}})$ and $\mathrm{AV}(\check{\mathbf{Z}}_{1d}^{\mathrm{CR}})$. Now, noting that

$$\begin{aligned}\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}^{-1}\,\mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}}) &= [\mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}}) + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})\}\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1}]^{-1} \\ &= \{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}\}^{-1},\end{aligned}$$

we can write

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

Similarly, $\mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \{\mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} + \mathrm{AV}(\tilde{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}\}^{-1}$ and $\mathrm{AV}(\check{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \{\mathrm{AV}(\check{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} + \mathrm{AV}(\check{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}\}^{-1}$. For $\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}$ and $\tilde{\mathbf{Z}}_{1d}^{\mathrm{CR}}$ we obtain

$$\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) - \mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{CR}}) = \{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})\}^{-1} + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}\}^{-1} - \{\mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} + \mathrm{AV}(\tilde{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}\}^{-1}.$$

Since the matrices $\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}})$ and $\mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{CR}})$ are non-negative definite and non-singular, they are positive definite. By theorem 1, $\mathrm{AV}(\tilde{\mathbf{Z}}_{id}^{\mathrm{R}}) \leqslant \mathrm{AV}(\hat{\mathbf{Z}}_{id}^{\mathrm{R}})$, which by a suitable result on inverses of such matrices (Harville (1997), page 434) implies that $\mathrm{AV}(\tilde{\mathbf{Z}}_{id}^{\mathrm{R}})^{-1} \geqslant \mathrm{AV}(\hat{\mathbf{Z}}_{id}^{\mathrm{R}})^{-1}$, so

$$\mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} + \mathrm{AV}(\tilde{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1} \geqslant \mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}.$$

By applying again the above-mentioned result we obtain

$$\{\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} + \mathrm{AV}(\hat{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}\}^{-1} \geqslant \{\mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{R}})^{-1} + \mathrm{AV}(\tilde{\mathbf{Z}}_{2d}^{\mathrm{R}})^{-1}\}^{-1}$$

and, hence,

$$\mathrm{AV}(\hat{\mathbf{Z}}_{1d}^{\mathrm{CR}}) \geqslant \mathrm{AV}(\tilde{\mathbf{Z}}_{1d}^{\mathrm{CR}}).$$

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

The proof of parts (b) and (c) is as in (a).

References

Deville, J. C. and Särndal, C. E. (1992) Calibration estimators in survey sampling. *J. Am. Statist. Ass.*, **87**, 367–382.

Elliott, M. R. and Davis, W. W. (2005) Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Appl. Statist.*, **54**, 595–609.

Estevao, V. M. and Särndal, C. E. (1999) The use of auxiliary information in design-based estimation for domains. *Surv. Methodol.*, **25**, 213–221.

Estevao, V. M. and Särndal, C. E. (2004) Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *J. Off. Statist.*, **20**, 645–669.

Harville, D. A. (1997) *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Hidiroglou, M. A. (2001) Double sampling. *Surv. Methodol.*, **27**, 143–154.

Hidiroglou, M. A. and Patak, Z. (2004) Domain estimation using linear regression. *Surv. Methodol.*, **30**, 67–78.

Marker, D. A. (2001) Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. *Surv. Methodol.*, **27**, 183–188.

Merkouris, T. (2004) Combining independent regression estimators from multiple surveys. *J. Am. Statist. Ass.*, **99**, 1131–1139.

Raghunathan, T. E. and Grizzle, J. E. (1995) A split questionnaire survey design. *J. Am. Statist. Ass.*, **90**, 54–63.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W. and Feuer, E. J. (2007) Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *J. Am. Statist. Ass.*, **102**, 474–486.

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

Renssen, R. H. and Nieuwenbroek, N. J. (1997) Aligning estimates for common variables in two or more sample surveys. *J. Am. Statist. Ass.*, **92**, 368–375.

Särndal, C. E. (1996) Efficient estimators with simple variance in unequal probability sampling. *J. Am. Statist. Ass.*, **91**, 1289–1300.

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

Statistics Netherlands (1998) Special issue: integration of household surveys; design advantages and methods. In *Netherlands Official Statistics*, vol. 13. Voorburg: Statistics Netherlands.

Webber, M., Latouche, M. and Rancourt, E. (2000) Harmonized calibration of income statistics. *Internal Document*. Statistics Canada, Ottawa.

Wu, C. (2004) Combining information from multiple surveys through the empirical likelihood method. *Can. J. Statist.*, **32**, 15–26.

Zieschang, K. D. (1990) Sample weighting methods and estimation of totals in the consumer expenditures survey. *J. Am. Statist. Ass.*, **85**, 986–1001.