Journal of Official Statistics, Vol. 39, No. 1, 2023, pp. 27–44, http://dx.doi.org/10.2478/JOS-2023-0002

# A Multivariate Regression Estimator of Levels and Change for Surveys Over Time

*Anne Konrad*[1] *and Yves Berger*[2]

Rotations are often used for panel surveys, where the observations remain in the sample for a predefined number of periods and then rotate out. The information of previous waves can be exploited to improve current estimates. We propose a multivariate regression estimator which captures all information available from both waves. By adding additional auxiliary variables describing the information of the rotational design, the proposed estimator captures the sample correlation between waves. It can be used for the estimation of levels and changes.

*Key words:* Generalized regression estimation; composite estimator; rotating samples.

## 1. Introduction

Repeated socioeconomic surveys are often the basis for evaluating changes and levels over time (e.g., Smith et al. 2003). Estimates are usually based on repeated or rotational surveys, which involve rotations, that is, units remain in a survey for a predefined number of waves and then are replaced by new sampled units (e.g., Gambino and Silva 2009; Kalton 2009; Eurostat 2012). There are different rotation schemes. In an in-for-x rotational design the units remain in the sample for x consecutive waves and then are replaced by new sampled units. In an x-(y)-z rotational design, the units remain in the sample for x consecutive waves, leave the sample for y waves and then return for z consecutive waves. Then they are dropped from the sample completely and replaced by new sampled units (e.g., Bonnéry et al. 2020, 170). We shall consider two waves, but the proposed approach can be extended to more than two waves (see Subsection 3.1).

Rotational designs give partially overlapping samples between waves. Thus, between two consecutive waves, we have units sampled at both waves (the overlapping units), units sampled only at the first wave (units that rotate out) and units sampled only at the second wave (units that rotate in). The sample information from the previous wave can be used to improve the current wave estimates. We expect to have more efficient estimates when the variables are correlated over time (Steel and McLaren 2008).

We propose a "multivariate generalised regression" (GREG) estimator that exploits the sample overlap between two waves, as well as the non-overlapping samples containing the units observed in only one of the waves. The proposed estimator includes "extended design variables" as additional auxiliary variables, which capture the sample correlation between the variables and the sample rotation. Thereby, it borrows strength from all available sample information on the variables of interest and the auxiliary variables from both waves. This may provide efficient change and levels estimates. Furthermore, the extended design

[1] Leibniz Institute for Educational Trajectories, Survey Statistical Mathods, 96047 Bamberg, Germany, Email: anne.konrad@lifbi.de
[2] University of Southampton, Social Statistics, SO17 1BJ, Southampton, UK, Email: Y.G.Berger@soton.ac.uk

© Statistics Sweden

variables capture the sample design information, such as stratification and unequal probabilities. The proposed estimator can be applied for rotational samples of any rotation scheme or for the simultaneous estimation of two or more consecutive waves; for example, impact evaluation surveys with a baseline and a post-intervention data collection.

The idea of including the sample information on variables of interest from previous waves is not new. Hansen et al. (1953) and Gurney and Daly (1965) introduced a class of composite estimators that exploit the sample overlap between two consecutive waves. The "modified regression estimator" of Singh et al. (1997) includes additional auxiliary variables based on the variables of interest from previous waves. However, for the new units that rotate in, the values of these additional variables are unknown and usually imputed. The control totals of the additional variables are also unknown and have to be estimated, which leads to a variance inflation of the current wave estimate. In contrast, the proposed estimator neither relies on imputation nor the estimation of unknown control totals.

The article is organised as follows. Section 2 introduces the basic framework on rotational surveys and GREG estimators. In Section 3, we derive the proposed multivariate GREG estimator and its properties. Asymptotic optimality and variance estimation is investigated in Section 4. Alternative estimators considered in the literature such as the modified regression estimator are discussed in Section 5. In Section 6, a Monte Carlo simulation study compares the proposed multivariate GREG estimator with the modified regression estimator. Section 7 summarises our results.

## 2. Rotation Design and Generalised Regression Estimator

Let $U = \{1, \ldots, i, \ldots, N\}$ be a population of $N$ units. Without loss of generality, we consider two waves ($t = 1$ and $t = 2$). The proposed estimator introduced in Section 3, will be extended to more than two waves in Subsection 3.1. We assume that the population units are the same in both waves. In practice, a change in a population can be handled by adjusting the weights and the sampling frame in the cases of birth, death and emigration.

Let $s_1$ be the first wave sample of size $n_1$ selected without-replacement from $U$. The first-order inclusion probability of unit $i$ for wave 1 and 2 are denoted respectively by $\pi_{i1} = Pr(i \in s_1)$ and $\pi_{i2} = Pr(i \in s_2)$, where $Pr(\cdot)$ denotes the probability with respect to the design. We assume that both sample sizes $n_1$ and $n_2$ are fixed. The common sample is $s_{12} = s_1 \cap s_2$, with a sample size $n_{12} = \#s_{12}$, where $0 \leq n_{12} \leq n_1$. We assume that $n_{12}$ is fixed, because this is a common feature of rotational designs. It is common practice to assume that the units that rotate out cannot rotate in; that is, $Pr\{i \in (s_2 \setminus s_1) | i \in s_1\} = 0$.

Stratification is often used in practice. We suppose that the population $U$ is stratified into $H$ strata $U_h$, such that $\cup_{h=1}^{H} U_h = U$. We assume that stratification is the same at both waves. Let $s_{t,h}$ be the $t$-th wave sample of size $n_{t,h}$ selected without-replacement from the population $U_h$, where $t = 1$ or 2. At wave $t$, the overall sample is $s_t = \cup_{h=1}^{H} s_{t,h}$ with a total sample size $n_t = \sum_{h=1}^{H} n_{t,h}$. We assume that we have a rotation within strata, that is, the common sample within $U_h$ is denoted by $s_{12,h} = s_{1,h} \cap s_{2,h}$, with a sample size $n_{12,h} = \#s_{12,h}$, fixed by design. The ratio $\theta_h = n_{12,h}/n_{1,h}$ is the fraction of the overlap within $U_h$. The quantities $\theta_h$ are allowed to vary between strata.

The objective is to estimate unknown population totals of a variable of interest $y$, for different waves. The total of wave $t$ is

$$\tau_{y_t} := \sum_{i \in U} y_{it},$$

where $y_{it}$ is the value of $y$ for a unit $i \in U$ at wave $t$. The Horvitz and Thompson (1952) estimator

$$\hat{\tau}_{y_t} := \sum_{i \in s_t} \frac{y_{it}}{\pi_{it}}$$

is a design-unbiased estimator of $\tau_{y_t}$. For estimation of a domain of interest, we impose $y_{it} = 0$ for the units $i$ outside the domain.

The efficiency can be improved by incorporating auxiliary information in the estimation process. A widely used model-assisted estimator based on auxiliary information, is the generalised regression (GREG) estimator (Hansen et al 1953; Cassel et al. 1977; Särndal, 1980; Isaki and Fuller 1982; Wright 1983). Let $\mathbf{x}_{it}$ be the $Q_t$-vector of auxiliary variables for a unit $i$ at wave $t$. Suppose that the vector of population totals $\boldsymbol{\tau}_{xt} = \sum_{i \in U} \mathbf{x}_{it}$ at wave $t$, is known from census, registers, or other reliable sources. The customary GREG estimator is defined by

$$\hat{\tau}^{g}_{y_t} := \hat{\tau}_{y_t} + \hat{\mathbf{B}}_t^{\top} (\tau_{x_t} - \hat{\tau}_{x_t}) \tag{1}$$

where

$$\hat{\boldsymbol{\tau}}_{x_t} := \sum_{i \in s_t} \frac{\mathbf{x}_{it}}{\pi_{it}}, \tag{2}$$

$$\hat{\mathbf{B}}_t := \left( \sum_{i \in s_t} \frac{\mathbf{x}_{it}\mathbf{x}_{it}^{\top}}{\pi_{it}} \right)^{-1} \sum_{i \in s_t} \frac{\mathbf{x}_{it} y_{it}}{\pi_{it}}. \tag{3}$$

The estimator (1) is motivated by the linear regression model

$$y_{it} = \mathbf{x}_{it}^{\top} \boldsymbol{\beta}_t + \varepsilon_{it}, \tag{4}$$

specifying the relationship between $y_{it}$ and $\mathbf{x}_{it}$, where $E(\varepsilon_{it}) = 0$, $V(\varepsilon_{it}) = \sigma^2$ and $\mathrm{E}(\varepsilon_{it}\varepsilon_{jt}) = 0$ for all $i \neq j$. If $V(\varepsilon_{it}) = v_{it}\sigma^2$, a weighted least-squares estimator can be used instead of Equation (3) to reflect heteroscedasticity. In order to simplify the notation, we shall assume $v_{it} = 1$. Nevertheless, when $v_{it} \neq 1$, they can be easily added to the regression coefficient (13) of the proposed estimator. The use of $v_{it}$ is more relevant for business surveys. Homoscedasticity ($v_{it} = 1$) is often assumed in household surveys (Steel and Clark 2007, 52).

The asymptotic design-unbiased estimator (1) does not depend on whether the model (4) holds or not. Its efficiency is driven by the predictive power of the model (cf. Särndal et al. 1992, 227, 239). Hereafter, we shall use a design-based approach, that is, the model (4) shall not be used for inference.

## 3. Proposed Multivariate Regression Estimator

Let us consider the "*combined sample*" defined by the set $s_b = s_1 \cup s_2$ comprising all units from both waves. The corresponding sample size of $s_b$ is denoted by $n_b = \#s_b = n_1 + n_2 - n_{12}$. Let the "*extended weighted variable of interest*" be defined by

$$\check{y}_{it} := \frac{y_{it}}{\pi_{it}} \delta\{i \in s_t\} \quad \text{for all } i \in s_b \quad \text{and } t = 1, 2, \tag{5}$$

where $\delta\{i \in s_t\} = 1$ if $i \in s_t$, and $\delta\{i \in s_t\} = 0$ otherwise. Note that $\check{y}_{i2} = 0$ for all units $i \in s_b \setminus s_2$ that rotates out. We also have $\check{y}_{i1} = 0$ for all units $i \in s_b \setminus s_1$ that rotates in. Figure 1 is a visual representation of two waves, with units on the horizontal axis and the two waves on the vertical axis.

The "*extended weighted auxiliary variables*" are defined by

$$\check{\mathbf{x}}_{it} := \frac{\mathbf{x}_{it}}{\pi_{it}} \delta\{i \in s_t\} \quad \text{for all } i \in s_b \quad \text{and } t = 1, 2. \tag{6}$$

The set of auxiliary variables used at $t = 1$ can be different from the one used at $t = 2$. The set of auxiliary variables can also be the same. This is usually the case for panel surveys.

Note that Equation (2) can also be re-written as $\hat{\tau}_{x_t} = \sum_{i \in s_b} \check{\mathbf{x}}_{it}$. We also consider "*extended design variables*" given by

$$\mathbf{z}_{it} := (z_{it,1}, \ldots, z_{it,h}, \ldots, z_{it,H})^\top \delta\{i \in s_t\} \quad \text{for all } i \in s_b \quad \text{and } t = 1, 2$$

with $z_{it,h} = 1$ if the unit $i$ belongs to stratum $h$ in wave $t$, and $z_{it,h} = 0$ otherwise. The vector $\mathbf{z}_{it}$ represents the sampling design information given by the stratification. The Hadamard product $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ will play a key role. It reveals the information induced by the rotation, because it identifies the units within the common sample. Indeed, the $h$-th component of $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ is equal to one if and only if the unit $i$ belongs to the common sample of strata $h$. This component equals zero if and only if the unit $i$ rotates in or out. Thus, $\mathbf{z}_{it}$ can be used to describe the sample information given by the rotation and the stratification.

It can be verified that

$$\sum_{i \in s_t} \mathbf{z}_{it} = \mathbf{n}_t \quad \text{and} \quad \sum_{i \in s_b} \mathbf{z}_{i1} \circ \mathbf{z}_{i2} = \mathbf{n}_{12}, \tag{7}$$

where

$$\mathbf{n}_t := (n_{t,1}, \ldots, n_{t,h}, \ldots, n_{t,H})^\top,$$
$$\mathbf{n}_{12} := (n_{12,1}, \ldots, n_{12,h}, \ldots, n_{12,H})^\top.$$

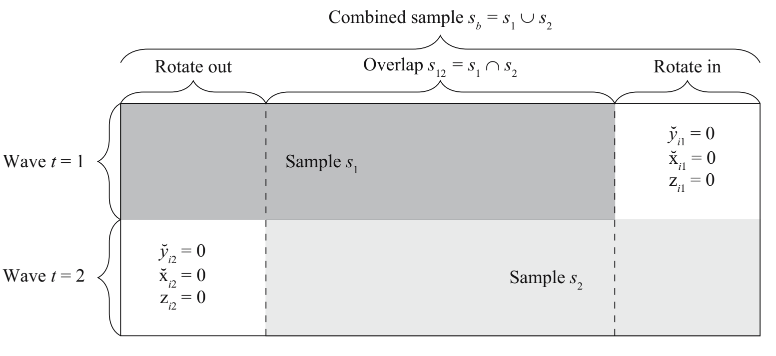


*Fig. 1. Visual representation of two waves. The vertical axis represents the two waves: $t = 1$ and $t = 2$. The horizontal axis represents the units of the combined sample $s_b = s_1 \cup s_2$. The sample $s_1$ and $s_2$ are given in two different gray scales:* *for the sample $s_1$* and for the sample $s_2$.

Equations (7) hold, because we have stratified design and we have a rotation within strata.

Let $\breve{\mathbf{y}}_i = (\breve{y}_{i1}, \breve{y}_{i2})^\top$ be the "*combined extended variable of interest*" of wave 1 and wave 2. We also pool together the extended weighted auxiliary variables and the extended design variables into a single vector $\gamma_i$ of dimension $(Q_1 + Q_2 + 3H)$, given by

$$\gamma_i := \{\breve{\mathbf{x}}_{i1}^\top, \breve{\mathbf{x}}_{i2}^\top, \mathbf{z}_{i1}^\top, \mathbf{z}_{i2}^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i2})^\top\}^\top. \tag{8}$$

This new auxiliary variable $\gamma_i$ contains the original auxiliary variables $\mathbf{x}_{it}$, the stratification variables $\mathbf{z}_{it}$ and the variables $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ which specify the rotation within strata.

Berger et al. (2003) proposed using the stratification variables as auxiliaries within a GREG estimator, when we have a single-stage stratified sampling designs. This has the merit of achieving asymptotic optimality. The resulting estimator is easy to implement and does not rely on joint-inclusion probabilities. The proposed multivariate GREG estimator (9) is based on a similar idea, except that we use the additional variables $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ to capture the rotation.

The proposed multivariate GREG estimator for the unknown vector $\boldsymbol{\tau}_y = (\boldsymbol{\tau}_{y_1}, \boldsymbol{\tau}_{y_2})^\top$ of totals, is defined by

$$\hat{\boldsymbol{\tau}}_y^{\text{greg}} := \hat{\boldsymbol{\tau}}_y + \hat{\mathbf{B}}_\gamma^\top (\boldsymbol{\tau}_\gamma - \hat{\boldsymbol{\tau}}_\gamma), \tag{9}$$

where

$$\hat{\boldsymbol{\tau}}_y := (\hat{\boldsymbol{\tau}}_{y_1}, \hat{\boldsymbol{\tau}}_{y_2})^\top, \tag{10}$$

$$\boldsymbol{\tau}_\gamma := (\boldsymbol{\tau}_{x_1}^\top, \boldsymbol{\tau}_{x_2}^\top, \mathbf{n}^\top)^\top, \tag{11}$$

$$\hat{\boldsymbol{\tau}}_\gamma := (\hat{\boldsymbol{\tau}}_{x_1}^\top, \hat{\boldsymbol{\tau}}_{x_2}^\top, \mathbf{n}^\top)^\top, \tag{12}$$

$$\hat{\mathbf{B}}_\gamma := \left( \sum_{i \in s_b} c_i \gamma_i \gamma_i^\top \right)^{-1} \sum_{i \in s_b} c_i \boldsymbol{\gamma}_i \breve{\mathbf{y}}_i^\top, \tag{13}$$

$$\mathbf{n} := (\mathbf{n}_1^\top, \mathbf{n}_2^\top, \mathbf{n}_{12}^\top)^\top, \tag{14}$$

$$c_i := 1 - Pr(i \in s_b). \tag{15}$$

The matrix (13) is a regression coefficient matrix of dimension $(Q_1 + Q_2 + 3H) \times 2$. We introduce the $c_i$ to achieve asymptotic optimality (see Section 4). Since $s_b = s_1 \cup s_2$, we have $Pr(i \in s_b) = \pi_{i1} + \pi_{i2} - Pr(i \in s_{12})$. Now, since $s_{12} = s_{12} \cap s_1$, $Pr(i \in s_{12}) = Pr(i \in s_1) Pr(i \in s_{12} | i \in s_1)$. Thus,

$$Pr(i \in s_b) = \pi_{i1} + \pi_{i2} - \pi_{i1} Pr(i \in s_{12} | i \in s_1). \tag{16}$$

The conditional probability $Pr(i \in s_{12} \mid i \in s_1)$ depends on the design and can be approximated by $\theta_h = n_{12,h}/n_{1,h}$ where $U_h \ni i$. Therefore, hereafter we shall use

$$c_i = 1 - \pi_{i1} - \pi_{i2} + \pi_{i1}\theta_h, \quad \text{where } h: \; U_h \ni i. \tag{17}$$

Exact computation of $Pr(i \in s_{12} \mid i \in s_1)$ is of little use. With large sampling fractions, the $c_i$ are less than 1 and can be interpreted as finite population corrections within Equation (13). They should not affect the consistency of Equation (9), because they are weights used only within Equation (13). Note that with negligible sampling fractions $c_i \approx 1$. The $c_i$ will be also used for variance estimation (see Equation (27)).

Because of nonresponse, we could have units within the overlapping sample, which are not available at both occasions. Re-weighting should be used to compensate for the missing observations. In this case, within Equations (5) and (6), the basic weights $\pi_{it}^{-1}$ should be replaced by weights that takes the missingness into account.

Theorem 1 gives an alternative expression for the proposed estimator which will be used to show its asymptotic optimality in Section 4.

**Theorem 1.** *An alternative expression for* $\hat{\boldsymbol{\tau}}_y^{greg}$ *is*

$$\hat{\boldsymbol{\tau}}_y^{greg} = \hat{\boldsymbol{\tau}}_y + \hat{\mathbf{B}}_x^\top (\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x), \tag{18}$$

where

$$\hat{\mathbf{B}}_x := (\breve{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z \breve{\mathbf{X}})^{-1} \breve{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z\, \breve{\mathbf{y}}, \tag{19}$$

$$\mathbf{M}_z := \mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{C}\mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{C}, \tag{20}$$

$$\begin{aligned} \breve{\mathbf{X}} &:= (\breve{\mathbf{x}}_1^\top, \ldots, \breve{\mathbf{x}}_{n_b}^\top)^\top, \\ \breve{\mathbf{y}} &:= (\breve{\mathbf{y}}_1^\top, \ldots, \breve{\mathbf{y}}_{n_b}^\top)^\top, \\ \mathbf{Z} &:= (\mathbf{z}_1^\top, \ldots, \mathbf{z}_{n_b}^\top)^\top, \\ \mathbf{C} &:= \mathrm{diag}\{c_1, \ldots, c_{n_b}\}, \end{aligned} \tag{21}$$

$$\begin{aligned} \breve{\mathbf{y}}_i &:= (\breve{y}_{i1}, \breve{y}_{i2})^\top, \\ \breve{\mathbf{x}}_i &:= (\breve{\mathbf{x}}_{i1}^\top, \breve{\mathbf{x}}_{i2}^\top)^\top, \\ \mathbf{z}_i &:= \{\mathbf{z}_{i1}^\top, \mathbf{z}_{i2}^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i2})^\top\}^\top, \end{aligned} \tag{22}$$

$$\begin{aligned} \boldsymbol{\tau}_x &:= (\boldsymbol{\tau}_{x_1}^\top, \boldsymbol{\tau}_{x_2}^\top)^\top, \\ \hat{\boldsymbol{\tau}}_x &:= (\hat{\boldsymbol{\tau}}_{x_1}^\top, \hat{\boldsymbol{\tau}}_{x_2}^\top)^\top \end{aligned} \tag{23}$$

*and* $\boldsymbol{I}$ *is the* $n_b \times n_b$ *identity matrix.*

The proof can be found in the Appendix (Section 8) and is based on the fact that the Horvitz-Thompson estimators of the totals of the design variables are equal to their population totals, that is, $\boldsymbol{\tau}_\gamma - \hat{\boldsymbol{\tau}}_\gamma = \{(\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x)^\top, \mathbf{0}^\top\}^\top$.

The underlying multivariate model that leads to Equation (18) is

$$\mathbf{y}_i = \mathbf{x}_i^\top \boldsymbol{\beta}_x + \varepsilon_i,$$

where $\mathbf{y}_i := (\pi_{i1}\breve{y}_{i1},\ \pi_{i2}\,\breve{y}_{i2})^\top$ and $\mathbf{x}_i := (\pi_{i1}\,\breve{\mathbf{x}}_{i1}^\top, \pi_{i2}\,\breve{\mathbf{x}}_{i2}^\top)^\top$. This model takes the correlation between waves into account, because variables of both waves are included within $\mathbf{y}_i$ and $\mathbf{x}_i$. Furthermore, $\hat{\boldsymbol{\tau}}_x$ contains the totals of both waves.

The proposed estimator borrows strength from both waves, by using both waves auxiliary variables. Furthermore, it takes the stratification into account, because of the extended design variables $\mathbf{z}_{i1}$ and $\mathbf{z}_{i2}$. In addition, the variable $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ exploits the rotation between $s_1$ and $s_2$ induced by the sample overlap. In contrast, the regression coefficient of the wave specific GREG estimator, given by Equation (3), does not involve design variables or information about the rotation. It does not take into account the correlation between the waves for the auxiliary variables and the variable of interest.

### *3.1. Extension to More than Two Waves*

The proposed estimator can be easily extended for more than two waves. Suppose we have three consecutive waves. At wave 2, the multivariate GREG estimator produces two estimates: $\hat{\tau}_{y_1}^{\text{greg}}$ for wave 1 and $\hat{\tau}_{y_2}^{\text{greg}}$ for wave 2, where $\hat{\tau}_{y_2}^{\text{greg}}$ borrows strength from the information of wave 1. At wave 3, we obtain a new estimate $\hat{\tau}_{y_2}^{\text{greg}(2)}$ for wave 2 and an estimate $\hat{\tau}_{y_3}^{\text{greg}}$ for wave 3. Therefore, we have two estimates for wave 2: $\hat{\tau}_{y_2}^{\text{greg}}$ and $\hat{\tau}_{y_2}^{\text{greg}(2)}$. In official statistics, due to the need for up-to-date information, the estimate $\hat{\tau}_{y_2}^{\text{greg}}$ is immediately published at wave 2. The second estimate $\hat{\tau}_{y_2}^{\text{greg}(2)}$ is not published and should not be viewed as a revised estimate for the second wave total. It is only used to produce $\hat{\tau}_{y_3}^{\text{greg}}$. Furthermore, there is no reason for $\hat{\tau}_{y_2}^{\text{greg}(2)}$ to be more precise than $\hat{\tau}_{y_2}^{\text{greg}}$, since both are based on the same controls and correlations. The estimates $\hat{\tau}_{y_2}^{\text{greg}}$ and $\hat{\tau}_{y_2}^{\text{greg}(2)}$ are not used as controls to produce $\hat{\tau}_{y_3}^{\text{greg}}$, as with the modified regression estimator (see Section 5).

The proposed estimator is flexible, because it can be also use to borrow strength over more than two waves. In this case, the dimension of the vectors $\hat{\boldsymbol{\tau}}_y^{\text{greg}}$ and $\check{\mathbf{y}}_i$ is the number of waves. The vectors $\check{\mathbf{y}}_i$ and $\check{\mathbf{x}}_i$ contain the variables of the waves considered. In this case, the vector (22) may need to include additional components depending on the design. For simplicity, we recommend to use $c_i = 1$ in this case.

For example, suppose we have three waves, the sample sizes of the overlapping sets between the three samples from the same stratum can be fixed by design; i.e $n_{12,h}, n_{23,h}, n_{13,h}$ and $n_{123,h}$ may be fixed, where $n_{tl,h}$ denotes the sample size of $s_{t,h} \cap s_{l,h}$ within stratum $U_h$. Here, $n_{123,h}$ is the sample size of $s_{1,h} \cap s_{2,h} \cap s_{3,h}$. This situation occurs when we use the customary rotation group method. In this case, we need to include within $\mathbf{z}_i$: (i) $\mathbf{z}_{i1} \circ \mathbf{z}_{i2}$ for the fixed sample size of $s_{1,h} \cap s_{2,h}$, (ii) $\mathbf{z}_{i2} \circ \mathbf{z}_{i3}$ for the fixed sample size of $s_{2,h} \cap s_{3,h}$, (iii) $\mathbf{z}_{i1} \circ \mathbf{z}_{i3}$ for the fixed sample size of $s_{1,h} \cap s_{3,h}$, (iv) $\mathbf{z}_{i1} \circ \mathbf{z}_{i2} \circ \mathbf{z}_{i3}$ for the fixed sample size of $s_{1,h} \cap s_{2,h} \cap s_{3,h}$; that is, the vectors (8) and (14) should be replaced respectively by

$$\gamma_i = \{\check{\mathbf{x}}_{i1}^\top, \check{\mathbf{x}}_{i2}^\top, \check{z}_{i1}^\top, \check{z}_{i2}^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i2})^\top, (\mathbf{z}_{i2} \circ \mathbf{z}_{i3})^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i3})^\top, (\mathbf{z}_{i1} \circ \mathbf{z}_{i2} \circ \mathbf{z}_{i3})^\top\}^\top,$$
$$\mathbf{n} = (\mathbf{n}_1^\top, \mathbf{n}_2^\top, \mathbf{n}_{12}^\top, \mathbf{n}_{23}^\top, \mathbf{n}_{13}^\top, \mathbf{n}_{123}^\top)^\top,$$

with $\mathbf{n}_{23} := (n_{23,1}, \ldots, n_{23,H})^\top$, $\mathbf{n}_{13} := (n_{13,1}, \ldots, n_{13,H})^\top$ and $\mathbf{n}_{123} := (n_{123,1}, \ldots, n_{123,H})^\top$.

## 4. Asymptotic Optimality and Variance Estimation

In this section, we show the asymptotic optimality when we have two waves; The asymptotic optimal GREG estimator (Montanari 1987) of the vector of totals $\boldsymbol{\tau}_y = (\tau_{y_1}, \tau_{y_2})^\top$ is

$$\hat{\boldsymbol{\tau}}_y^{\text{opt}} := \hat{\boldsymbol{\tau}}_y + \hat{\mathbf{B}}_{\text{opt}}^\top (\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x), \tag{24}$$

where

$$\hat{\mathbf{B}}_{\text{opt}}^\top := \hat{\mathbf{V}}(\hat{\boldsymbol{\tau}}_x)^{-1} \widehat{\mathbf{Cov}}(\hat{\boldsymbol{\tau}}_x, \hat{\boldsymbol{\tau}}_y). \tag{25}$$

See Guandalini and Tillé (2017, 3) for more details. By using the Horvitz and Thompson (1952) variance and covariance estimators, the expression (25) reduces to

$$\hat{\mathbf{B}}_{\text{opt}} = (\breve{\mathbf{X}}^\top \boldsymbol{\Delta} \breve{\mathbf{X}})^{-1} \breve{\mathbf{X}}^\top \boldsymbol{\Delta} \breve{\mathbf{y}}, \tag{26}$$

where

$$\boldsymbol{\Delta} := \{(\pi_{ij} - \pi_i \pi_j)\pi_{ij}^{-1}; i,j \in s_b\}.$$

Here, $\pi_{ij} = Pr(i, j \in s_b)$ denotes the joint-inclusion probability of units $i$ and $j$ for the sample $s_b$. These are different from the joint probabilities of $s_1$ and $s_2$, because $\pi_{ij}$ takes the rotation into account. Since the probabilities $\pi_{ij}$ are unknown, we propose to use the asymptotic approximation of Hájek (1964), based on the assumption that the rotation design is asymptotically rejective according to the design constraints (7). This approximation is given by $\boldsymbol{\Delta} \approx \mathbf{C}\mathbf{M}_z$, where $\mathbf{C}$ and $\mathbf{M}_z$ are defined respectively by Equations (20) and (21) (Hájek 1981 chap. 14; Berger et al. 2003; and Deville and Tillé 2005). Now, by replacing this approximation of $\boldsymbol{\Delta}$ within Equation (26), we obtain Equation (19). Thus, the proposed estimator $\hat{\boldsymbol{\tau}}_y^{\text{greg}}$ is indeed optimal asymptotically.

A variance of the estimator (9) can be derived, based on principle that the variance under Poisson sampling of the regression estimator (9) based on the auxiliary and design variables, is asymptotically the same as the variance of the regression estimator (18) under a rejective design (Hájek 1964; Berger 2004) with the design constraints (7). Thus, the variance estimator of (9), assuming that $s_b$ is a Poisson sample with inclusion probabilities (16), is given by the variance-covariance matrix

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\tau}}_{\mathbf{y}}^{\text{greg}}) := (\mathbf{M}_{\boldsymbol{\Gamma}} \breve{\mathbf{y}})^\top \mathbf{C}\, \mathbf{M}_{\boldsymbol{\Gamma}} \breve{\mathbf{y}}, \tag{27}$$

where

$$\mathbf{M}_\Gamma := \mathbf{I} - \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top \mathbf{C} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^\top \mathbf{C},$$
$$\boldsymbol{\Gamma} := (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n)^\top.$$

Note that (27) is a residual variance as in Särndal et al. (1992, 235), because $\mathbf{M}_{\boldsymbol{\Gamma}} \breve{\mathbf{y}}$ are residuals. Note that the variance estimator takes the stratification into account, because the information about the strata is included within $\mathbf{M}_{\boldsymbol{\Gamma}}$. However, if within Equations (5) and (6), the basic weights $\pi_{it}^{-1}$ are substituted by weights which take the missingness into account, the variance estimator (27) may be biased, because nonresponse is not accounted for.

## 5. Alternative Approaches

Composite estimators also use the information from previous waves. Hansen et al. (1953) introduced the $K$-composite estimator for levels and change between two waves. The $AK$-composite estimator (Gurney and Daly 1965) takes the difference between the common sample $s_{12}$ and the unmatched sample $s_2$ into account. The optimal choice of the weighting factors $A$ and $K$, within the $AK$-composite estimator, depends on the variables of interest (Kumar et al. 1983). This dependency may result in an inconsistency, in the sense that the sub-group total estimates may not add up to the overall total (Gambino et al. 2001, 66).

Singh (1996) and Singh et al. (1997) introduced the modified regression estimator, abbreviated MR hereafter. The idea is to extend the auxiliary variables in the current wave by an additional artificial auxiliary variable, which contains the information on the variable of interest from the previous wave. The definition of this variable depends on whether the primary interest lies on levels or change. If the main focus lies on levels, the artificial variable refers to the variable of interest $y_{i1}$ from the previous wave. However, due to the rotation, $y_{i1}$ is only known for $i \in s_{12}$. Singh (1996) suggested to use mean imputation for the unknown values for the units $i \in s_2 \setminus s_{12}$. Thus, in this case, the artificial variable is

$$\tilde{x}_{i2}^{\mathrm{MR1}} := \begin{cases} y_{i1} & \text{for} \quad i \in s_{12} \\ \hat{\mu}_{y1} & \text{for} \quad i \in s_2 \setminus s_{12}, \end{cases} \tag{28}$$

where $\hat{\mu}_{y_1}$ is an estimator for the mean of $y_1$. The control total of the variable (28) is unknown and can be estimated by $N\hat{\mu}_{y_1}$ (Fuller and Rao 2001, 47). Hence, the modified regression estimator for $\tau_{y_2} = \sum_{i \in U_2} y_{i2}$ is given by

$$\hat{\tau}_{y_2}^{\mathrm{MR1}} := \hat{\tau}_{y2} + \hat{\mathbf{B}}_{x\tilde{x}}^{\top}(\tilde{\tau}_{x\tilde{x}} - \hat{\tau}_{x\tilde{x}}), \tag{29}$$

with

$$\hat{\mathbf{B}}_{x\tilde{x}} := (\mathbf{B}_{x_2}^{\top}, \hat{B}_{\tilde{x}_2})^{\top},$$

$$\tilde{\boldsymbol{\tau}}_{x\tilde{x}} := (\boldsymbol{\tau}_{x_2}^{\top}, N\hat{\mu}_{y_1})^{\top},$$
$$\hat{\boldsymbol{\tau}}_{x\tilde{x}} := (\hat{\boldsymbol{\tau}}_{x_2}^{\top}, \hat{\tau}_{\tilde{x}_2})^{\top}.$$

If the primary interest is to estimate a change, the artificial variable refers to the variable of interest $y_{i2}$ from the current wave. The variable recommended by Singh (1996) and Singh et al. (1997) is

$$\tilde{x}_{i2}^{\mathrm{MR2}} := \begin{cases} y_{i2} + \dfrac{n_2}{n_{12}}(y_{i1} - y_{i2}) & \text{for } i \in s_{12}, \\ y_{i2} & \text{for } i \in s_2 \setminus s_{12}. \end{cases} \tag{30}$$

The MR2 estimator may suffer from a drift in levels estimates over a long period (Gambino et al. 2001, 65; Fuller and Rao 2001, 50). In order to overcome this problem, Fuller and Rao (2001) introduced the regression composite estimator (RC) given by

$$\tilde{x}_{i2}^{\mathrm{RC}} := (1 - \alpha)\tilde{x}_{i2}^{\mathrm{MR1}} + \alpha\tilde{x}_{i2}^{\mathrm{MR2}}, \tag{31}$$

where $\alpha \in [0, 1]$ is a tune-in parameter which reflects the importance given to levels or change estimates. The advantage of the regression composite estimator compared with MR1 and MR2 is the fact that it is a compromise between levels and change estimation. An alternative estimator could be based on Definitions (28) and (30). However, the increased number of auxiliaries and control totals may lead to a distortion in the final weights (Gambino et al. 2001, 65).

Singh et al. (2001) suggested a jackknife variance estimator that takes the estimation of the control totals into account. Indeed, ignoring the additional source of randomness would lead to an underestimation of the true variance. Berger et al. (2009) proposed a linearised variance estimator that takes the estimation of the controls into account.

The optimal BLUE estimator is based on a time series of the variable of interest (Yansaneh and Fuller 1998; Bell 2001; Australian Bureau of Statistics 2007). This estimator requires that the variances and covariances of the rotation group estimates are known (Bell 2001, 56). If they were substituted by their estimates, it is no longer guaranteed that the BLUE estimator is optimal. Bonnéry et al. (2020) showed that the BLUE with an estimated variance-covariance matrix is less efficient than the composite estimators. Some disadvantages are discussed in Fuller (1990) and Steel and McLaren (2009). Since the BLUE estimator is based on a time series, it is less comparable with the proposed estimator and the modified estimators, which are both based on regression estimation.

## 6. Simulation Study

We consider three waves ($t = 0, 1, 2$), because the estimators (9) and (29) at wave $t = 1$, require the sample information from wave $t = 0$. The results are reported for levels at waves $t = 1$ and $t = 2$, and changes between waves $t = 1$ and $t = 2$.

Consider $N$ population values of $y_{it}$ and $x_{it}$ ($t = 1, 2, 3$) generated from a multivariate normal distribution; that is,

$$(y_{i0}, y_{i1}, y_{i2}, x_{i0}, x_{i1}, x_{i2})^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Here, $\boldsymbol{\Sigma}$ denotes a covariance matrix with an heterogeneous exchangeable structure, that is,

$$\boldsymbol{\Sigma} := \mathrm{diag}(\sigma)\{\rho \mathbf{J}_6 + (1-\rho)\mathbf{I}_6\}\mathrm{diag}(\sigma).$$

where diag($\sigma$) is the diagonal matrix with $\sigma = (\sigma_{y0,}\ \sigma_{y1,}\ \sigma_{y2,}\ \sigma_{x0,}\ \sigma_{x1,}\ \sigma_{x2})^\top$ as its diagonal. The matrices $\mathbf{J}_6$ is $6 \times 6$ matrix of ones and $\mathbf{I}_6$ is the $6 \times 6$ identity matrix. Thus, the correlations $cor(y_{it},\ y_{it'}) = cor(x_{it},\ x_{it'}) = cor(y_{it},\ x_{it'}) = \rho$, with $t \neq t'$. Let $\sigma_{y_0} = 10$, $\sigma_{y_1} = 15$, $\sigma_{y_2} = 20$, $\sigma_{x_0} = 30$, $\sigma_{x_1} = 40$ and $\sigma_{x_2} = 50$. The correlations considered are $\rho = 0.1, 0.5$ and $0.9$. Two values for the vector $\boldsymbol{\mu} = (\mu_{y0}, \mu_{y1}, \mu_{y2}, \mu_{\mathbf{x}0}, \mu_{x1}, \mu_{x2})^\top$ are used:

$$\boldsymbol{\mu}_I := (59, 60, 61, 99, 100, 101)^\top,$$
$$\boldsymbol{\mu}_{II} := (40, 60, 80, 100, 150, 200)^\top,$$

that is, we have a small change with $\boldsymbol{\mu}_I$ and a large change with $\boldsymbol{\mu}_{II}$.

For each wave $t$, we have stratified samples of size $n_t = 1{,}000$. We consider four strata formed by the quantile classes of the population distribution of $y_{i1} + y_{i2}$. The same fraction of common samples between waves is used within strata, that is, $\theta_h = \theta = n_{12}/n_1 = n_{01}/n_0$, where $\theta = 0.25$, $\theta = 0.5$ or $\theta = 0.75$. Rotation groups sampling is implemented. Within each strata $U_h$, $q$ units are randomly allocated into $P$ rotation groups of same size $p = \lfloor q/P \rfloor$. The sample $s_{0,h}$ contains the units of the first $\lfloor n_h p^{-1} \rfloor$ groups. At wave $t = 1$, we obtain the sample $s_{1,h}$ by rotating out the first group and replacing it by the ($\lfloor n_h p^{-1} \rfloor + 1$)-th group. At wave $t = 2$, the second group rotates out and ($\lfloor n_h p^{-1} \rfloor + 2$)-th group rotates in. For $\theta = 0.25$, we use $q = 625$ and $P = 10$. With $\theta = 0.5$, we use $q = 400$ and $P = 4$ and with $\theta = 0.75$, we set $q = 300$ and $P = 6$. We consider 1,000 iterations.

In the first simulation setup, we consider equal allocation for all strata with $n_{t,h} = 250$ and $N = 100{,}000$. Thus, the inclusion probabilities are the same across the strata and the sampling fractions are small. In the second simulation setup, unequal probabilities with large sampling fractions are used. Consider $n_{1,h} = 50$, $n_{2,h} = 200$, $n_{3,h} = 350$, $n_{4,h} = 400$ and $N = 4{,}000$. The resulting within strata inclusion probabilities are 0.05, 0.2, 0.35 and

0.4. In the second simulation setup, the population size is $N = 4{,}000$, to allow for large sampling fractions.

The estimators considered are the proposed multivariate regression estimator (9) (PROP), the customary regression estimator (1) (GREG) and the modified estimator (29) with (28) as auxiliaries (MR1) and with (30) as auxiliaries (MR2). For MR1 and MR2, we use $\hat{\tau}_{y_0}^{\text{greg}}$ as the estimated control total of the previous wave $t = 1$.

In order to explore the efficiency of point estimates, we compare the empirical '*relative root mean squared errors*' (RRMSE). Let $\hat{\tau}_r$ be an estimate for the $r$-th iteration with $r = 1, \ldots, 1{,}000$. The RRMSE is defined as

$$\text{RRMSE}(\hat{\tau}) := \frac{1}{|\tau|}\left\{\frac{1}{1{,}000}\sum_{r=1}^{1{,}000}(\hat{\tau}_r - \tau)^2\right\}^{1/2},$$

where $\tau$ denotes the population total.

The RRMSE $\times$ 100% for different values of $\rho$, $g$ and $\theta$, are reported in Tables 1 and 2. For Table 1, we have equal strata sizes with the same inclusion probabilities across strata and small sampling fractions. For Table 2, the inclusion probabilities differs between strata and some sampling fractions are large. The proposed PROP estimator outperforms GREG and MR1 under all scenarios. For all estimators under consideration, the RRMSE decreases with the correlation $\rho$ between the variables.

We observe slightly smaller RRMSE for MR1 with $\theta = 0.75$, when $\rho = 0.1$, because the higher the overlap, the less values have to be imputed. The amount of overlap $\theta$ has little

*Table 1. Equal strata sizes. equal probabilities and small sampling fractions. RRMSE X 100% of levels estimates under different scenarios for 1,000 iterations.*

| $\rho$ | $\boldsymbol{\mu}$ | $\theta$ | GREG | | PROP | | MR1 | |
|---|---|---|---|---|---|---|---|---|
| | | | $t = 1$ | $t = 2$ | $t = 1$ | $t = 2$ | $t = 1$ | $t = 2$ |
| 0.1 | $\boldsymbol{\mu}_I$ | 0.25 | 1.21 | 1.49 | 0.62 | 0.69 | 1.00 | 1.14 |
| | | 0.50 | 1.26 | 1.44 | 0.63 | 0.68 | 0.99 | 1.12 |
| | | 0.75 | 1.24 | 1.40 | 0.62 | 0.67 | 0.92 | 1.06 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 1.21 | 1.12 | 0.62 | 0.53 | 1.37 | 1.27 |
| | | 0.50 | 1.26 | 1.08 | 0.63 | 0.52 | 1.29 | 1.15 |
| | | 0.75 | 1.24 | 1.06 | 0.62 | 0.51 | 1.16 | 1.04 |
| 0.5 | $\boldsymbol{\mu}_I$ | 0.50 | 1.03 | 1.26 | 0.48 | 0.54 | 0.78 | 0.84 |
| | | 0.50 | 0.99 | 1.27 | 0.49 | 0.58 | 0.80 | 0.90 |
| | | 0.75 | 0.99 | 1.25 | 0.50 | 0.56 | 0.80 | 0.90 |
| | $\boldsymbol{\mu}_{II}$ | 0.50 | 1.03 | 0.92 | 0.48 | 0.41 | 1.06 | 0.83 |
| | | 0.50 | 0.99 | 0.93 | 0.49 | 0.44 | 1.09 | 0.89 |
| | | 0.75 | 0.99 | 0.91 | 0.50 | 0.43 | 1.09 | 0.91 |
| 0.9 | $\boldsymbol{\mu}_I$ | 0.25 | 0.49 | 0.60 | 0.28 | 0.35 | 0.39 | 0.46 |
| | | 0.50 | 0.49 | 0.61 | 0.27 | 0.35 | 0.40 | 0.45 |
| | | 0.75 | 0.48 | 0.60 | 0.28 | 0.33 | 0.37 | 0.43 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 0.49 | 0.43 | 0.28 | 0.27 | 0.51 | 0.37 |
| | | 0.50 | 0.49 | 0.43 | 0.27 | 0.27 | 0.53 | 0.38 |
| | | 0.75 | 0.48 | 0.42 | 0.28 | 0.25 | 0.53 | 0.41 |

*Table 2. Unequal strata sizes. unequal probabilities and some large sampling fractions. RRMSE X 100% of levels estimates under different scenarios for 1,000 iterations.*

| $\rho$ | $\boldsymbol{\mu}$ | $\theta$ | GREG | | PROP | | MR1 | |
|---|---|---|---|---|---|---|---|---|
| | | | $t=1$ | $t=2$ | $t=1$ | $t=2$ | $t=1$ | $t=2$ |
| 0.1 | $\boldsymbol{\mu}_I$ | 0.25 | 1.51 | 1.61 | 0.84 | 0.89 | 1.26 | 1.21 |
| | | 0.50 | 1.51 | 1.58 | 0.87 | 0.88 | 1.24 | 1.26 |
| | | 0.75 | 1.53 | 1.58 | 0.87 | 0.90 | 1.23 | 1.20 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 1.51 | 1.26 | 0.84 | 0.68 | 1.67 | 1.29 |
| | | 0.50 | 1.51 | 1.24 | 0.87 | 0.67 | 1.60 | 1.27 |
| | | 0.75 | 1.53 | 1.24 | 0.87 | 0.68 | 1.53 | 1.14 |
| 0.5 | $\boldsymbol{\mu}_I$ | 0.25 | 1.29 | 1.51 | 0.70 | 0.77 | 0.98 | 0.98 |
| | | 0.50 | 1.32 | 1.42 | 0.71 | 0.74 | 1.03 | 0.99 |
| | | 0.75 | 1.26 | 1.48 | 0.69 | 0.74 | 1.04 | 1.04 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 1.29 | 1.14 | 0.70 | 0.59 | 1.30 | 0.93 |
| | | 0.50 | 1.32 | 1.07 | 0.71 | 0.56 | 1.38 | 0.98 |
| | | 0.75 | 1.26 | 1.11 | 0.69 | 0.56 | 1.42 | 1.07 |
| 0.9 | $\boldsymbol{\mu}_I$ | 0.25 | 0.72 | 0.83 | 0.38 | 0.50 | 0.53 | 0.61 |
| | | 0.50 | 0.70 | 0.85 | 0.40 | 0.49 | 0.53 | 0.60 |
| | | 0.75 | 0.71 | 0.84 | 0.40 | 0.48 | 0.53 | 0.59 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 0.72 | 0.59 | 0.38 | 0.38 | 0.69 | 0.50 |
| | | 0.50 | 0.70 | 0.60 | 0.40 | 0.37 | 0.73 | 0.51 |
| | | 0.75 | 0.71 | 0.59 | 0.39 | 0.37 | 0.78 | 0.58 |

impact on the precision of the proposed estimator. However, for MR1, we observe some slight differences in the RRMSE between different values for $\theta$. For small correlation ($\rho = 0.1$), we indeed have a larger RRMSE for MR1 with $\theta = 0.25$. For larger correlation, the differences are negligible for MR1. With MR1, we notice differences between the RRMSE of $\hat{\tau}_{y_1}$ for small ($\boldsymbol{\mu}_I$) and large changes ($\boldsymbol{\mu}_{II}$). There are hardly any differences for the proposed estimator. These observations are the same for Tables 1 and 2, except that the RRMSE are higher for all estimators in case of unequal strata sizes.

The RRMSE of the proposed estimator does not seem to be affected by the amount of overlap $\theta$, because we can see from the expression (18) that the precision is driven by the correlations between the variables of interest and the auxiliary information for both waves, which is not affected by $\theta$. This can also be seen from the variance (27), where the residuals $\mathbf{M}_{\Gamma}\check{\mathbf{y}}$ do not depend on $\theta$. The information about the rotation is implicitly included within the vector $\mathbf{z}_i$ given by Equation (22), and used for the weights within the regression coefficient (19) (see Equation (20)). These weights ensure efficiency (see Section 4). On the other hand, the precision of MR1 is related to $\theta$, because $\theta$ has an impact on the precision of the control totals with MR1. With the proposed method, we use different control totals unaffected by $\theta$.

Let $\triangle = \tau_{y_2} - \tau_{y_1}$ be the change between waves $t = 1$ and $t = 2$. We propose estimating $\triangle$ by $\hat{\triangle} = \hat{\tau}_{y_2} - \hat{\tau}_{y_1}$, where $\hat{\tau}_{y_1}$ and $\hat{\tau}_{y_2}$ are the corresponding cross-sectional estimators. Tables 3 and 4 give the RRMSE $\times$ 100% of the estimators of changes, for equal and unequal strata sizes. As expected, the RRMSE decreases with $\rho$. The proposed estimator PROP significantly outperforms GREG and MR2. The efficiency gain compared with MR2

*Table 3. Equal strata sizes. equal probabilities and small sampling fractions. RRMSE × 100% of change estimates under different scenarios for 1,000 iterations.*

| $\rho$ | $\boldsymbol{\mu}$ | $\theta$ | GREG | PROP | MR2 |
|---|---|---|---|---|---|
| 0.1 | $\boldsymbol{\mu}_I$ | 0.25 | 120.28 | 60.60 | 100.83 |
| | | 0.50 | 125.23 | 67.39 | 90.90 |
| | | 0.75 | 118.08 | 69.03 | 82.45 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 5.86 | 2.98 | 5.40 |
| | | 0.50 | 6.10 | 3.31 | 5.05 |
| | | 0.75 | 5.76 | 3.39 | 4.37 |
| 0.5 | $\boldsymbol{\mu}_I$ | 0.25 | 98.04 | 46.96 | 79.61 |
| | | 0.50 | 99.73 | 51.67 | 71.26 |
| | | 0.75 | 94.29 | 52.56 | 64.99 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 4.76 | 2.34 | 4.42 |
| | | 0.50 | 4.85 | 2.57 | 4.12 |
| | | 0.75 | 4.59 | 2.61 | 3.60 |
| 0.9 | $\boldsymbol{\mu}_I$ | 0.25 | 46.14 | 27.00 | 37.37 |
| | | 0.50 | 45.38 | 26.22 | 32.06 |
| | | 0.75 | 43.62 | 26.43 | 28.95 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 2.23 | 1.36 | 2.15 |
| | | 0.50 | 2.19 | 1.32 | 2.01 |
| | | 0.75 | 2.11 | 1.32 | 1.74 |

*Table 4. Unequal strata sizes. unequal probabilities and some large sampling fractions. RRMSE × 100% of change estimates under different scenarios for 1,000 iterations.*

| $\rho$ | $\boldsymbol{\mu}$ | $\theta$ | GREG | PROP | MR2 |
|---|---|---|---|---|---|
| 0.1 | $\boldsymbol{\mu}_I$ | 0.25 | 140.96 | 79.27 | 123.09 |
| | | 0.50 | 138.27 | 86.79 | 109.39 |
| | | 0.75 | 144.61 | 93.57 | 110.40 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 6.90 | 3.83 | 6.43 |
| | | 0.50 | 6.76 | 4.19 | 5.79 |
| | | 0.75 | 7.06 | 4.51 | 5.75 |
| 0.5 | $\boldsymbol{\mu}_I$ | 0.25 | 114.65 | 61.44 | 95.13 |
| | | 0.50 | 113.51 | 61.60 | 81.51 |
| | | 0.75 | 109.77 | 62.43 | 74.85 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 5.95 | 3.20 | 5.48 |
| | | 0.50 | 5.89 | 3.21 | 4.92 |
| | | 0.75 | 5.69 | 3.25 | 4.37 |
| 0.9 | $\boldsymbol{\mu}_I$ | 0.25 | 64.40 | 37.30 | 48.98 |
| | | 0.50 | 61.67 | 36.67 | 41.70 |
| | | 0.75 | 60.04 | 35.11 | 36.51 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 3.16 | 1.91 | 2.96 |
| | | 0.50 | 3.03 | 1.88 | 2.61 |
| | | 0.75 | 2.95 | 1.80 | 2.20 |

ranges from 5% to 53%. Since the relative RMSE is considered, it is not surprising to observe larger RRMSE for a small change ($\boldsymbol{\mu}_I$). The RRMSE of MR2 decreases with $\theta$. In contrast, the RRMSE of PROP increases slightly with $\theta$ except for large vales of $\rho$.

Table 5 shows the relative bias (RB) of the variance estimator (27) for PROP. The RB is defined by

$$\mathrm{RB}\{\hat{V}(\hat{\tau})\} := V(\hat{\tau})^{-1}\left\{\frac{1}{1{,}000}\sum_{r=1}^{1{,}000}\hat{V}(\hat{\tau}_r) - V(\hat{\tau})\right\},$$

where

$$V(\hat{\tau}) := \frac{1}{1{,}000}\sum_{r=1}^{1{,}000}(\hat{\tau}_r - \tau)^2.$$

Here, $\hat{\tau}_r$ and $\hat{V}(\hat{\tau}_r)$ are respectively the point and variance estimate for the $r$-th iteration. The RB are within an acceptable range. We observe larger RB for $\hat{\tau}_{y_2}^{\text{greg}}$ when $\rho = 0.9$ and $\theta = 0.75$, because the variance is small in this case.

The biases of the variance estimates in the case of unequal strata sizes is larger than the biases of equal strata sizes. The reason is the small sample size for two strata in the unequal strata size scenario. The residuals of the smallest strata vary much more and thus have a larger contribution towards the variance than the residuals of the large strata. The negative bias can also be caused by small sample sizes, because the Taylor linearization method has a tendency to underestimate the true variance in this case (Särndal et al. 1992, 176).

*Table 5. RB%100 of variance estimates for the proposed estimator under different scenarios for 1,000 iterations.*

| $\rho$ | $\boldsymbol{\mu}$ | $\theta$ | Equal strata sizes | | Unequal strata sizes | |
|---|---|---|---|---|---|---|
| | | | $t=1$ | $t=2$ | $t=1$ | $t=2$ |
| 0.1 | $\boldsymbol{\mu}_I$ | 0.25 | 1.5 | −7.4 | −16.3 | −15.3 |
| | | 0.50 | −3.4 | −5.1 | −18.1 | −12.1 |
| | | 0.75 | 1.5 | 0.0 | −15.4 | −11.9 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 1.5 | −7.4 | −16.4 | −15.3 |
| | | 0.50 | −3.4 | −5.1 | −18.1 | −12.1 |
| | | 0.75 | 1.5 | 0.0 | −15.2 | −11.9 |
| 0.5 | $\boldsymbol{\mu}_I$ | 0.25 | 5.7 | 4.9 | −21.0 | −19.3 |
| | | 0.50 | 1.8 | −8.8 | −18.3 | −12.9 |
| | | 0.75 | −0.0 | −1.4 | −12.2 | −7.6 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | 5.6 | 4.9 | −20.8 | −19.3 |
| | | 0.50 | 1.8 | −8.8 | −18.3 | −12.9 |
| | | 0.75 | −0.0 | −1.4 | −11.5 | −7.6 |
| 0.9 | $\boldsymbol{\mu}_I$ | 0.25 | −2.7 | 0.3 | −16.1 | −22.0 |
| | | 0.50 | 4.0 | −1.4 | −18.5 | −18.7 |
| | | 0.75 | −2.0 | 10.1 | −15.1 | −13.7 |
| | $\boldsymbol{\mu}_{II}$ | 0.25 | −2.7 | 0.3 | −15.7 | −22.0 |
| | | 0.50 | 4.0 | −1.4 | −18.5 | −18.7 |
| | | 0.75 | −1.9 | 10.1 | −14.3 | −13.7 |

## 7. Conclusion

We propose a multivariate GREG estimator for estimation of levels and changes. It has the advantage of involving the information from both waves, and takes into account the correlations between the variables of interest and the auxiliaries within and between the waves. Additionally, it also takes the sampling design into account, in terms of stratification, rotation and sampling fractions.

The simulation study shows that the proposed estimator may outperform its competitors, in particular with respect to change estimates. Nevertheless, the advantages of the proposed estimator over the modified estimator are manifold. It does not require any imputation and does not suffer from a drift, unlike the composite estimator. It can be easily implemented using existing statistical software. The variance estimator is simpler than the variance estimator of composite estimators, because neither estimated totals nor imputation is required. It also takes the auxiliary variables and the variables of interest from both waves into account.

Nonresponse and panel attrition are important issues with repeated surveys. It is beyond the scope to tackle these problems fully. Previous wave imputation can be used for the auxiliary variables $\breve{\mathbf{x}}_{it}$ which suffer from attrition. Re-weighting could be used to compensate for nonresponse and panel attrition for the variable of interest. In this case the new weight should replace the basic weights $1/\pi_{it}$ within Equations (5) and (6). In this case, $s_t$ would be the sample of respondents at wave $t$. The proposed estimator (9) can be directly used in this case. It is approximately unbiased, as long as a proper re-weighting technique has been used. However, in this case, the vectors $\mathbf{n}_t$ and $\mathbf{n}_{12}$ are random. Consequently, we may lose the asymptotic optimality, because the asymptotic approximation of Hájek (1964) for the joint-inclusion probabilities are based on fixed $\mathbf{n}_t$ and $\mathbf{n}_{12}$. The variance estimator (27) should be used cautiously, because it does not incorporate nonresponse adjustments. A possible solution would be to incorporate the re-weighting variables within $\breve{\mathbf{x}}_{it}$ and use $\breve{\mathbf{x}}_{it}$ within Equations (9) and (27). It would be useful to investigate this idea further.

## 8. Appendix

**Proof of Theorem 1:**
Since $\boldsymbol{\gamma}_i = (\breve{\mathbf{x}}_i^\top, \mathbf{z}_i^\top)^\top$, we have

$$\left(\sum_{i\in s_b} c_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^\top\right)^{-1} = \left\{\begin{pmatrix}\breve{\mathbf{X}}\\ \mathbf{Z}\end{pmatrix}^\top \mathbf{C}\begin{pmatrix}\breve{\mathbf{X}}\\ \mathbf{Z}\end{pmatrix}\right\}^{-1} = \begin{pmatrix}\boldsymbol{\Gamma}_{xx} & \boldsymbol{\Gamma}_{xz}\\ \boldsymbol{\Gamma}_{xz}^\top & \boldsymbol{\Gamma}_{zz}\end{pmatrix},$$

$$\sum_{i\in s_b} c_i \boldsymbol{\gamma}_i \breve{\mathbf{y}}_i^\top = \begin{pmatrix}\breve{\mathbf{X}}\\ \mathbf{Z}\end{pmatrix}^\top \mathbf{C}\,\breve{\mathbf{y}},$$

where

$$\boldsymbol{\Gamma}_{xx} = (\breve{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z \breve{\mathbf{X}})^{-1},$$

$$\boldsymbol{\Gamma}_{zz} = (\mathbf{Z}^\top \mathbf{C}\mathbf{M}_x \mathbf{Z})^{-1},$$

$$\boldsymbol{\Gamma}_{xz} = -\boldsymbol{\Gamma}_{xx}\breve{\mathbf{X}}^\top \mathbf{C}\mathbf{Z}(\mathbf{Z}^\top \mathbf{C}\mathbf{Z})^{-1}$$

and $\mathbf{M}_x$ is defined by

$$\mathbf{M}_x = \mathbf{I} - \check{\mathbf{X}}(\check{\mathbf{X}}^\top \mathbf{C}\check{\mathbf{X}})^{-1}\check{\mathbf{X}}^\top \mathbf{C}. \tag{32}$$

Now, we have

$$\hat{\mathbf{B}}_\gamma = \begin{pmatrix} \hat{\mathbf{B}}_x \\ \boldsymbol{\Gamma}_{xz}^\top \check{\mathbf{X}}^\top \mathbf{C}\check{\mathbf{y}} + \boldsymbol{\Gamma}_{zz}\mathbf{Z}^\top \mathbf{C}\check{\mathbf{y}} \end{pmatrix}, \tag{33}$$

because

$$\begin{aligned} \boldsymbol{\Gamma}_{xx}\check{\mathbf{X}}^\top \mathbf{C}\check{\mathbf{y}} + \boldsymbol{\Gamma}_{xz}\mathbf{Z}^\top \mathbf{C}\check{\mathbf{y}} &= \boldsymbol{\Gamma}_{xx}\check{\mathbf{X}}^\top \mathbf{C}\{\mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top \mathbf{C}\mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{C}\}\check{\mathbf{y}} \\ &= \boldsymbol{\Gamma}_{xx}\check{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z\check{\mathbf{y}} \\ &= (\check{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z\check{\mathbf{X}})^{-1}\check{\mathbf{X}}^\top \mathbf{C}\mathbf{M}_z\check{\mathbf{y}} \\ &= \hat{\mathbf{B}}_x. \end{aligned}$$

Finally, Equations (11) and (12) imply that $\boldsymbol{\tau}_\gamma - \hat{\boldsymbol{\tau}}_\gamma = \{(\boldsymbol{\tau}_x - \hat{\boldsymbol{\tau}}_x)^\top, 0\}^\top$ Thus, by using Equation (33), we obtain Equation (18).

## 9. References


Australian Bureau of Statistics. 2007. *Forthcoming changes to labour force statistics:* 6292.0, Australian Bureau of Statistics, Canberra, Australia. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/mf/6292.0 (accessed July 2022).

Bell, P. 2001. "Comparison of alternative labour force survey estimators." *Survey Methodology* 27: 53–63. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5854-eng.pdf.

Berger, Y.G. 2004. "Variance estimation for measures of change in probability sampling." *Canadian Journal of Statistics* 32: 451–467. DOI: https://doi.org/10.2307/3316027.

Berger, Y.G., J.F. Muñoz, and E. Rancourt. 2009. "Variance estimation of survey estimates calibrated on estimated control totals – an application to the extended regression estimator and the regression composite estimator." *Computational Statistics and Data Analysis* 53: 2596–2604. DOI: https://doi.org/10.1016/j.csda.2008.12.011.

Berger, Y.G., M.E.H. Tirari, and Y. Tillé. 2003. "Towards optimal regression estimation in sample surveys." *Australian & New Zealand Journal of Statistics* 45: 319–329. DOI: https://doi.org/10.1111/1467-842X.00286.

Bonnéry, D., Y. Cheng, and P. Lahiri. 2020. "An evaluation of design-based properties of different composite estimators." *Statistics in Transition New Series* 21: 166–190. DOI: https://doi.org/10.21307/stattrans-2020-037.

Cassel, C.-M., C. Särndal, and J. Wretman. 1977. *Foundations of inference in survey sampling*. New York: Wiley.

Deville, J.C., and Y. Tillé. 2005. "Variance approximation under balanced sampling." *Journal of Statistical Planning and Inference* 128: 569–591. DOI: https://doi.org/10.1016/j.jspi.2003.11.011.

Eurostat. 2012. *European union statistics on income and living conditions* (EU-SILC). Available at: https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions (accessed July 2022).

Fuller, W. 1990. "Analysis of repeated surveys." *Survey Methodology* 16: 167–180. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199000214537.

Fuller, W.A., and J. Rao. 2001. "A regression composite estimator with application to the canadian labour force survey." *Survey Methodology* 27: 45–52. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5853-eng.pdf.

Gambino, J., B. Kennedy, and M.P. Singh. 2001. "Regression composite estimation for the canadian labour force survey: Evaluation and implementation." *Survey Methodology* 27: 65–74. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5855-eng.pdf.

Gambino, J.G., and P.L.N. Silva. 2009. "Sampling and estimation in house-hold surveys." In *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao. 29A of Handbook of Statistics: 407–439. Amsterdam: Elsevier.

Guandalini, A., and Y. Tillé. 2017. "Design-based estimators calibrated on estimated totals from multiple surveys." *International Statistical Review* 85: 250–269. DOI: https://doi.org/10.1111/insr.12160.

Gurney, M.A., and J.F. Daly. 1965. "A multivariate approach to estimation in periodic sample surveys." In *Proceedings of the Section of Survey and Research Methods*, American Statistical Association, September 1965: 242–257, Philadelphia USA. Available at: http://www.asasrms.org/Proceedings/index.html? (accessed February 2023).

Hájek, J. 1964. "Asymptotic theory of rejective sampling with varying probabilities from a finite population." *The Annals of Mathematical Statistics* 35: 1491–1523. DOI: https://doi.org/10.1214/aoms/1177700375.

Hájek, J. 1981. *Sampling from a Finite Population*. New York: Marcel Dekker.

Hansen, M., Hurwitz, W., and W. Madow. 1953. *Sample survey methods and theory*, vol. I and II. New York: Wiley.

Horvitz, D.G., and D.J. Thompson. 1952. "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association* 47: 663–685. DOI: https://doi.org/10.1080/01621459.1952.10483446.

Isaki, C.T., and W.A. Fuller. 1982. "Survey design under the regression superpopulation model." *Journal of the American Statistical Association* 77: 89–96. DOI: https://doi.org/10.1080/01621459.1982.10477770.

Kalton, G. 2009. "Design for surveys over time." In *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao, 29A of *Handbook of Statistics*: 89–108. Amsterdam: Elsevier.

Kumar, S., and H. Lee. 1983. "Evaluation of composite estimation for the canadian labor force survey." *Survey Methodology* 9: 403–408. Availabe at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198300214342.

Montanari, G. 1987. "Post sampling efficient qr-prediction in large sample survey." *International Statistical Review* 55: 191–202. DOI: https://doi.org/10.2307/1403195.

Särndal, C.E. 1980. "On $\pi$-inverse weighting versus best linear unbiased weighting in probability sampling." *Biometrika* 67: 639–650. DOI: https://doi.org/10.1093/biomet/67.3.639.

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

Singh, A.C. 1996. "Combining information in survey sampling by modified regression." . In Proceedings of the Section on Survey Research Methods, American Statistical Association. August 1996: 120–129. Chicago, USA. Available at: http://www.asasrms.org/Proceedings/index.html? (accessed February 2023).

Singh, A.C., B. Kennedy, and S. Wu. 2001. "Regression composite estimation for the canadian labour force survey with a rotating panel design." *Survey Methodology* 27: 33–44. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2001001/article/5852-eng.pdf.

Singh, A.C., B. Kennedy, S. Wu, and F. Brisebois. 1997. "Composite estimation for the canadian labour force survey." In *Proceedings of the Survey Research Methods Section*, American Statistical Association, August 1997: 300–3005, Anaheim, USA. Available at: http://www.asasrms.org/Proceedings/index.html? (accessed February 2023).

Smith, P., M. Pont, and T. Jones. 2003. "Developments in business survey methodology in the office for national statistics, 1994–2000." *Journal of the Royal Statistical Society.* 52: 257–295. DOI: https://doi.org/10.1111/1467-9884.03571.

Steel, D., and R. Clark. 2007. "Person-level and household-level regression estimation in household surveys." *Surveys Methodology* 33: 55–60. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20070019852.

Steel, D., and C. McLaren. 2008. *Design and analysis of repeated surveys*. Centre for Statistical and Survey Methodology, University of Wollongong: 11-08. Available at: http://ro.uow.edu.au/cgi/viewcontent.cgi?article = 1009&context = cssmwp (accessed July 2022).

Steel, D., and C. McLaren. 2009. "Design and analysis of surveys repeated over time." In *Handbook of statistics*, edited by C.R. Rao. 289–313 Elsevier.

Wright, R.L. 1983. "Finite population sampling with multivariate auxiliary information." *Journal of the American Statistical Association* 78: 879–884. DOI: https://doi.org/10.1080/01621459.1983.10477035.

Yansaneh, I.S., and W.A. Fuller. 1998. "Optimal recursive estimation for repeated surveys." *Survey Methodology* 24: 31–40. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1998001/article/3907-eng.pdf.

Received December 2021
Revised April 2021
Accepted August 2022