ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 21, Number 1, June 2017
Available online at http://acutm.math.ut.ee

Effect of auxiliary information in data collection and estimation stage

Kaur Lumiste

Abstract. Responsive design is a newly emerged view focusing on reducing the effects of non-response by monitoring and intervening the data collection process. Informative measures that use auxiliary information are used to guide the data collection process. Aspiration to a well representative set of respondents is currently done through balancing – means of auxiliary variables have to be equal in the sample and the set of respondents. Auxiliary variables are later used in the estimation stage to improve the estimates, but assume that more auxiliary variables are available in the estimation stage. The auxiliary vector is split by variables (a) used in monitoring and estimation, and (b) only used in the estimation stage. Explicit terms of calibration weights and response propensities are developed and useful properties of those terms are proved. Theoretical results and two emerging strategies are tested in simulations.

1. Introduction

High levels of non-response have become an almost unavoidable part of every survey, leading to biased results and questionable inference. National statistics offices around the world are forced to find new ways of countering non-response bias in their surveys. There is extensive literature on how to reduce non-response bias of estimates in the estimation stage of the survey, but corrective actions can and should be taken earlier – in planning and data collection phases. With the recent development of computer-assisted methods of data collection, survey researchers now have the ability to continuously monitor the collection of survey data and process data (paradata). This creates the opportunity for a possible change of emphasis during the course of data collection to achieve more precise, less biased estimates, and

Received October 25, 2016.
2010 *Mathematics Subject Classification.* 62D05.
Key words and phrases. Non-response bias; balance; imbalance; monitoring response.
http://dx.doi.org/10.12697/ACUTM.2017.21.08

to improve survey cost efficiency. Such surveys are labeled as "responsive designs" in the ground breaking article [3].

There is a growing number of research being done on responsive survey designs, an extensive overview is given in [16], but some examples include [6], [14], [15] and [8], where the goal is to get a well representative set of respondents through planning and appropriate intervention in the data collection process. One option is to monitor data collection with more advanced response quality measures than response rate, like R-indicators [15] or fraction of missing information (FMI) [1]. In the present paper we aspire to a "representative" set of respondents through reducing the difference of auxiliary variable means between the response and the sample. A scalar indicator developed in [8] is used to measure this difference.

An equivalent approach is to estimate response propensities and use their variation as a measure of imbalance and lower variation signals for less imbalance. These propensities are estimated using auxiliary variables and assess balance with an imbalance measure discussed in [8]. This approach gives effective methods for interventions during data collection, for example the threshold method and the fixed proportion method proposed in [11].

Despite all efforts with monitoring data collection, a perfectly balanced response is nearly impossible to achieve and non-response bias hard to evade. So auxiliary variables are used also in the estimation stage to reduce non-response bias, usually in constructing calibrated weights, but they need not be the same variables used in monitoring. Assume that we have access to additional calibration variables after data collection. It can be in the form of paradata from the data collection process, like the number of contact attempts or interviewer notes on approaching respondents [5], or data from registers becomes available that was not present when the sample was drawn. Another reason why the lists of monitoring and calibration stage variables may differ is that auxiliary variables for the estimation may be updated versions of the same variables available in the data collection.

Is the effect of additional explanation power of new auxiliary variables affected by balancing? Should more emphasize be put on acquiring more auxiliary variables in the estimation stage or balancing the response during data collection? Which would have a larger effect on the bias of the final estimates?

Answering these questions is important for budgeting, since monitoring data collection and acquiring extra explaining power means extra costs and/or extra effort. The process of monitoring response and acting accordingly often means additional work for the survey agency and interviewers doing fieldwork. Gathering paradata means developing universal methodology, and train interviewers to follow that methodology to avoid large measurement error. Accessing different national registries might be free of charge, especially for national or research institutions, but gained information may

have incomplete or error data, meaning additional effort for imputation or data cleaning.

We consider an auxiliary vector divided into two parts by usage – auxiliary variables used for (a) both for monitoring in data collection stage and for calibration in the estimation stage, and (b) new variables added to the auxiliary vector in the estimation stage. A similar situation was touched in [11], but here a more focused research is presented and calibration weights explicitly showing the effect of both parts of the auxiliary vector are developed. The results are interpreted in the light of monitoring data collection with the first part of the auxiliary vector. We also develop the propensity expression explicitly showing two parts of the auxiliary information. Based on this, a measure for the balancing effect on additional auxiliary information is given.

The article is arranged in 3 parts: Sections 2–3 give the necessary notation and concepts of imbalance and its measurement. Sections 4–6 bring in the situation of split auxiliary variables, calibration weights and response propensities are developed for this case. Properties and special cases of the newfound results are studied. In Section 7 a simulation study is described, where theoretical results are confirmed and aspects, complementary to theoretical part, discussed.

2. Preliminaries

Let $U = \{1, 2, \ldots, N\}$ denote a finite population of N units and a probability sample s of size n is selected to estimate some characteristics of U. The probability sampling design generates for an element k a known inclusion probability, $P(k \in s) = \pi_k > 0$, and a corresponding design weight $d_k = 1/\pi_k$. In case of non-response, data can only be collected from a subset r within the sample, $r \subset s \subset U$, and the values y_k of the study variable y are recorded for the units $k \in r$ only. It is assumed that there is access to auxiliary information on unit level, i.e., the vector of J auxiliary variables $\mathbf{x}_k = (\mathbf{x}_{1k}, \mathbf{x}_{2k}, \ldots, \mathbf{x}_{Jk})'$ is known for every element $k \in U$ (or minimally for every element $k \in s$) and the auxiliary variable vector satisfies

$$\boldsymbol{\mu}'\mathbf{x}_k = 1, \forall k \in U, \text{ for some constant vector } \boldsymbol{\mu}. \tag{1}$$

It is not a major restriction as most vectors $\mathbf{x}_k$ of importance in practice are of this kind. For example, for a numerical auxiliary variable x_k take $\mathbf{x}_k = (1, x_k)'$ and $\boldsymbol{\mu} = (1, 0)'$ satisfies the requirement. When $\mathbf{x}_k = (0, 0, \ldots, 1, \ldots, 0)'$, as in coding a set of mutually exclusive and exhaustive categories of units, then $\boldsymbol{\mu} = (1, 1, \ldots, 1)'$ satisfies the requirement.

The conducted survey's objective is to estimate the population total $Y = \sum_U y_k$ using the collected data y_k, $k \in r$, and auxiliary information $\mathbf{x}_k \in s$ (here $\sum_A$ denotes a sum over all the units k in set A). The basic design

unbiased estimator of Y from a full sample is $\hat{Y}_{FUL} = \sum_s d_k y_k$, the Horwitz–Thompson (HT) estimator, but in the presence of non-response this cannot be computed. In such a situation one can use the simple expansion estimator

$$\hat{Y}_{EXP} = \hat{N}\bar{y}_r, \tag{2}$$

where $\hat{N} = \sum_s d_k$ estimates the population size and $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ is the design weighted mean of the study variable in r. But this estimator is often considerably biased, so a more widely used method is to calibrate on the auxiliary vector $\mathbf{x}_k$:

$$\hat{Y}_{CAL} = \sum_r d_k g_k y_k, \tag{3}$$

where

$$g_k = \left(\sum_s d_k \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \tag{4}$$

are the calibration weights (g-weights for short). Notice that the weights g_k satisfy the sample level calibration requirement $\sum_r d_k g_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$, where $\sum_s d_k \mathbf{x}_k$ are unbiased estimates for population totals $\sum_U \mathbf{x}_k$.

3. Measuring balance

The concept of balance has been often used in statistical literature with reference to an equality of means of certain variables for two sets of units, where one is the subset of the other. For example balanced sampling aims to give a random sample so that the means of a set of auxiliary variables are equal (or approximately equal) within the sample and the population. One such sampling method is the Cube Method [2]. Here we look at measuring balance of auxiliary variables in the response set and the sample, and the desirable, but often unreachable, balance of the study variable.

3.1. Measuring balance of the response set. Auxiliary information can be used already in the data collection phase to realize a well-balanced set of respondents. The concept and measuring of lack of balance, i.e., imbalance, has been thoroughly discussed in [8].

With the given auxiliary vector $\mathbf{x}$, means can be calculated for the response, $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k / \sum_r d_k$, and for the sample, $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k / \sum_s d_k$. If $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, then the response set is said to be perfectly balanced on the given $\mathbf{x}$-vector. In practice this is usually not the case and the J-dimensional mean difference $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ signals drift from perfect balance. A univariate indicator of imbalance is defined as

$$IMB = P^2 \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \Sigma_s^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right),$$

where $\Sigma_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k$ is a $J \times J$ weighting matrix assumed non-singular, and $P = \sum_r d_k / \sum_s d_k$. IMB is a value with $0 < IMB < P(1-P)$

and can be calculated at any point during data collection and situates r in relation to s in respect of the chosen $\mathbf{x}$-vector.

An alternative way to define an imbalance measure is by measuring the variance of response propensities f_k:

$$IMB_{alt} = \mathrm{var}_s(f) = \frac{\sum_s d_k \left(f_k - \bar{f}_s\right)^2}{\sum_s d_k},$$

where

$$f_k = \left(\sum_r d_k \mathbf{x}_k\right)' \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \tag{5}$$

are response propensities (also mentioned as monitoring weights, f-weights). Notice that $\sum_r d_k \mathbf{x}_k = \sum_s d_k f_k \mathbf{x}_k$ and more intensively,

$$\bar{f}_s = \sum_s d_k f_k / \sum_s d_k = \sum_r d_k / \sum_s d_k = P.$$

This approach is related to the R-indicator (R for representativity) from [15], where f-weights (5) would be linear estimates of their conditional response probabilities.

It turns out that $IMB = IMB_{alt}$ as shown in, for example, [10], but both are needed for a clearer interpretation in latter sections. Due to equality the notation IMB will be used for both approaches.

3.2. Balance in the study variable. Full sample mean of the y-variable, $\bar{y}_s = \sum_s d_k y_k / d_k$, is not available under non-response. Although unknown, it is essentially unbiased for the population mean, while the computable mean of the response set $\bar{y}_r = \sum_r d_k y_k / d_k$, often is not. The degree of relationship between the study variable y and the chosen $\mathbf{x}$-variables is a major influence on non-response bias. The linear regression coefficient vectors are $\mathbf{b}_r$ (computable) for the response set fit and $\mathbf{b}_s$ (conceptually defined) for the full sample fit, where

$$\begin{aligned} \mathbf{b}_r &= \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_r d_k \mathbf{x}_k y_k\right), \\ \mathbf{b}_s &= \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_s d_k \mathbf{x}_k y_k\right). \end{aligned}$$

Using the regression coefficient vectors $\mathbf{b}_r$ and $\mathbf{b}_s$, imbalance for the study variable $\bar{y}_r - \bar{y}_s$ can be split into two terms:

$$\bar{y}_r - \bar{y}_s = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r + \left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s. \tag{6}$$

Notice that $\bar{\mathbf{x}}_r' \mathbf{b}_r = \bar{y}_r$ and $\bar{\mathbf{x}}_s' \mathbf{b}_s = \bar{y}_s$ due to the property $\boldsymbol{\mu}' \mathbf{x}_k = 1$, for all k and $\bar{\mathbf{x}}_s' \mathbf{b}_r = \hat{Y}_{CAL} / \sum_s d_k$. This split illustrates two undesirable differences associated with non-response, $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ caused by imbalance and $\mathbf{b}_r - \mathbf{b}_s$

caused by inconsistent regression. Another viewpoint is from the aspect of calibration – the first term $(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \mathbf{b}_r$ shows how much calibration adjusts the simple, but often biased response mean $\bar{y}_r$, the second term $(\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s$ shows how much calibration deviates from the unbiased full sample mean $\bar{y}_s$. This can be better illustrated if both sides of (6) are multiplied by $\hat{N} = \sum_s d_k$:

$$\hat{N}\left(\bar{y}_r - \bar{y}_s\right) = \left(\hat{Y}_{EXP} - \hat{Y}_{CAL}\right) + \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right). \tag{7}$$

The effect of balancing is clear with the first term in (6), but effect on the second term is not so straightforward. The recent article [9] showed that balancing on the auxiliary vector increases the chance to get $\hat{Y}_{CAL}$ closer to $\hat{Y}_{FUL}$ and therefore a smaller difference in the second term of (7).

4. Splitting the auxiliary vector

Thus far the same auxiliary vector is used throughout the entire survey process, now the auxiliary vector used in the estimation stage is split up to two parts. Let the auxiliary vector comprise of

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{Mk} \\ \mathbf{x}_{Ck} \end{pmatrix}, \tag{8}$$

where $\mathbf{x}_{Mk} : p \times 1$ is an auxiliary vector used earlier for monitoring response and $\mathbf{x}_{Ck} : q \times 1$ is an auxiliary vector of extra set of variables that are later added to compute the calibration estimator in the estimation stage. Assume that $\mathbf{x}_{Mk}$ satisfies

$$\boldsymbol{\mu}_M' \mathbf{x}_{Mk} = 1, \forall k \in U \tag{9}$$

for some constant vector $\boldsymbol{\mu}_M$. Note that if (9) is satisfied, then, for example,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_M \\ \mathbf{0}_q \end{pmatrix}$$

satisfies (1) for any $\mathbf{x}_{Ck}$, where $\mathbf{0}_q : q \times 1$ is a vector of zeros.

A split up of auxiliary vector is discussed in [12, pp. 53–56], but only in the context of post-weighting and the split is done by the level of detail of auxiliary information – for some auxiliary variables, values are known for the whole population and for the remaining, only for sampled elements.

4.1. Calibration weights. Calibration weights g_k, defined in (4), for the long vector in (8) can be expressed using partitioned matrices (the indexes M and C indicate whether the auxiliary vector $\mathbf{x}_{Mk}$ or $\mathbf{x}_{Ck}$ is used, and the following r or s indicates the summation set of units)

$$g_k = \begin{pmatrix} \mathbf{T}_{Ms} \\ \mathbf{T}_{Cs} \end{pmatrix}' \begin{pmatrix} \mathbf{T}_{MMr} & \mathbf{T}_{MCr} \\ \mathbf{T}_{CMr} & \mathbf{T}_{CCr} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_{Mk} \\ \mathbf{x}_{Ck} \end{pmatrix},$$

where $\mathbf{T}_{Ms} = \sum_s d_k \mathbf{x}_{Mk}$, $\mathbf{T}_{Cs} = \sum_s d_k \mathbf{x}_{Ck}$ are design weighted totals of M- and C-information auxiliary variables in the response set and

$$\begin{array}{rclcrcl} \mathbf{T}_{MMr} & = & \sum_r d_k \mathbf{x}_{Mk}\mathbf{x}'_{Mk}; & & \mathbf{T}_{MCr} & = & \sum_r d_k \mathbf{x}_{Mk}\mathbf{x}'_{Ck}; \\ \mathbf{T}_{CCr} & = & \sum_r d_k \mathbf{x}_{Ck}\mathbf{x}'_{Ck}; & & \mathbf{T}_{CMr} & = & \mathbf{T}'_{MCr}. \end{array}$$

Proposition 1. *When the auxiliary vector* (8) *is split into two, calibration weights g_k can be separated into two terms:*

$$g_k = g_{Mk} + h_k,$$

where

$$g_{Mk} = \mathbf{T}'_{Ms}\mathbf{T}^{-1}_{MMr}\mathbf{x}_{Mk}, \tag{10}$$

$$h_k = \mathbf{T}'_{\varepsilon s}\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\varepsilon_k. \tag{11}$$

and $\mathbf{T}_{\varepsilon s} = \sum_s d_k \varepsilon_k$ is a sum of the residuals $\varepsilon_k = \mathbf{x}_{Ck} - \mathbf{B}'_r\mathbf{x}_{Mk}$ over the sample units, $\mathbf{B}_r = \mathbf{T}^{-1}_{MMr}\mathbf{T}_{MCr}$ is a matrix of coefficients, and $\mathbf{T}_{\varepsilon\varepsilon r} = \sum_r d_k \varepsilon_k \varepsilon'_k$ is a q-dimensional square matrix of residuals and assumed non-singular.

The proof is presented in the Appendix.

When distinguishing auxiliary information according to usage, one can see that the calibration weights g_k can be divided into two, one part dependent only on auxiliary information available or used for monitoring response, the other uses both M- and C-information – residual information from C-information not linearly explained by M-information. The g_M and h-weights are uncorrelated over the response set as the following result shows.

Proposition 2. *The h-weights in a response set r have the following properties:*

$$\bar{h}_r = \frac{\sum_r d_k h_k}{\sum_r d_k} = 0, \tag{12}$$

$$var_r(h) = \frac{\sum_r d_k \left(h_k - \bar{h}_r\right)^2}{\sum_r d_k} = P^{-1}\bar{h}_s, \tag{13}$$

$$cov_r(g_M, h) = \frac{\sum_r d_k \left(g_{Mk} - \bar{g}_{Mr}\right)\left(h_k - \bar{h}_r\right)}{\sum_r d_k} = 0, \tag{14}$$

$$cov_r(f_M, h) = \frac{\sum_r d_k \left(f_{Mk} - \bar{f}_{Mr}\right)\left(h_k - \bar{h}_r\right)}{\sum_r d_k} = 0, \tag{15}$$

where $\bar{h}_s = \sum_s d_k h_k / \sum_s d_k$ and

$$\bar{g}_{Mr} = \sum_r d_k g_{Mk} / \sum_r d_k, \ \bar{f}_{Mr} = \sum_r d_k f_{Mk} / \sum_r d_k, \ f_{Mk} = \mathbf{T}'_{Mr}\mathbf{T}^{-1}_{MMs}\mathbf{x}_{Mk}$$

are response propensities using only M-information and

$$\mathbf{T}_{Mr} = \sum_r d_k \mathbf{x}_{Mk}, \quad \mathbf{T}_{MMs} = \sum_s d_k \mathbf{x}_{Mk}\mathbf{x}'_{Mk}.$$

Proof. The design weighted mean of h-weights over the response set r gives

$$\bar{h}_r = \mathbf{T}'_{\varepsilon s}\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\frac{\sum_r d_k\boldsymbol{\varepsilon}_k}{\sum_r d_k}$$

and since $\sum_r d_k\boldsymbol{\varepsilon}_k$ is a vector of zeros, we arrive at (12).

Using (12) one can show that

$$\begin{aligned}\mathrm{var}_r\,(h) = \frac{\sum_r d_k h_k^2}{\sum_r d_k} &= \mathbf{T}'_{\varepsilon s}\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\frac{\sum_r d_k\boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}'_k}{\sum_r d_k}\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\mathbf{T}_{\varepsilon s}\\ &= \frac{1}{\sum_r d_k}\mathbf{T}'_{\varepsilon s}\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\mathbf{T}_{\varepsilon s}\\ &= \frac{\sum_s d_k}{\sum_r d_k}\frac{\sum_s d_k\boldsymbol{\varepsilon}'_k\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\mathbf{T}_{\varepsilon s}}{\sum_s d_k}\\ &= P^{-1}\bar{h}_s.\end{aligned}$$

With (12) the covariance of g and h-weights in r simplifies to

$$\mathrm{cov}_r(g_M, h) = \frac{\sum_r d_k g_{Mk}h_k}{\sum_r d_k}.$$

Expanding the weight terms, the residuals $\boldsymbol{\varepsilon}_k$, and $\mathbf{B}_r$ gives

$$\begin{aligned}\mathrm{cov}_r(g_M, h) &= \frac{1}{\sum_r d_k}\mathbf{T}'_{Ms}\mathbf{T}^{-1}_{MMr}\left[\sum_r d_k\mathbf{x}_{Mk}\left(\mathbf{x}_{Ck} - \mathbf{B}'_r\mathbf{x}_{Mk}\right)'\right]\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\mathbf{T}_{\varepsilon s}\\ &= \frac{1}{\sum_r d_k}\mathbf{T}'_{Ms}\mathbf{T}^{-1}_{MMr}\left[\mathbf{T}_{MCr} - \mathbf{T}_{MMr}\mathbf{T}^{-1}_{MMr}\mathbf{T}_{MCr}\right]\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\mathbf{T}_{\varepsilon s}\\ &= 0\end{aligned}$$

since $\mathbf{T}_{MCr} - \mathbf{T}_{MMr}\mathbf{T}^{-1}_{MMr}\mathbf{T}_{MCr}$ is a $p \times q$ matrix of zeros.

Analogously

$$\begin{aligned}\mathrm{cov}_r(f_M, h) &= \frac{1}{\sum_r d_k}\mathbf{T}'_{Mr}\mathbf{T}^{-1}_{MMs}\left[\sum_r d_k\mathbf{x}_{Mk}\left(\mathbf{x}_{Ck} - \mathbf{B}'_r\mathbf{x}_{Mk}\right)'\right]\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\mathbf{T}_{\varepsilon s}\\ &= \frac{1}{\sum_r d_k}\mathbf{T}'_{Mr}\mathbf{T}^{-1}_{MMs}\left[\mathbf{T}_{MCr} - \mathbf{T}_{MMr}\mathbf{T}^{-1}_{MMr}\mathbf{T}_{MCr}\right]\mathbf{T}^{-1}_{\varepsilon\varepsilon r}\mathbf{T}_{\varepsilon s}\\ &= 0.\end{aligned}$$

□

Remark. Note that the residuals $\boldsymbol{\varepsilon}_k$ in $\mathbf{T}_{\varepsilon s}$ in (11) are summed over the sample and in fact $\mathbf{T}_{\varepsilon s} = \sum_s d_k\boldsymbol{\varepsilon}_k = \sum_{s-r} d_k\boldsymbol{\varepsilon}_k$, since $\sum_r d_k\boldsymbol{\varepsilon}_k = \mathbf{0}_q$.

When there is perfect balance on the monitoring variable, i.e., $\bar{\mathbf{x}}_{Mr} = \bar{\mathbf{x}}_{Ms}$, then by using (9), calibration weights g_{Mk} and h-weights simplify to $g_{Mk} = P^{-1}$ for all k, and $h_k = \hat{N}\left(\bar{\mathbf{x}}_{Cs} - \bar{\mathbf{x}}_{Cr}\right)' \mathbf{T}^{-1}_{\varepsilon\varepsilon r}\boldsymbol{\varepsilon}_k$, because $\bar{\boldsymbol{\varepsilon}}_s = \sum_s d_k\boldsymbol{\varepsilon}_k / \sum_s d_k = \left(\bar{\mathbf{x}}_{Cs} - \bar{\mathbf{x}}_{Cr}\right)$. The calibration estimator in this case is the simple expansion estimator of the study variable (2) plus a correction term dependent on the balance of C-information and the regression of $\mathbf{x}_{Ck}$ on $\mathbf{x}_{Mk}$:

$$\begin{aligned}\hat{Y}_{CAL} &= P^{-1}\sum_r d_k y_k + \hat{N}\left(\bar{\mathbf{x}}_{Cs} - \bar{\mathbf{x}}_{Cr}\right)' \mathbf{T}^{-1}_{\varepsilon\varepsilon r}\sum_r d_k y_k\boldsymbol{\varepsilon}_k \\ &= \hat{Y}_{EXP} + \hat{N}\left(\bar{\mathbf{x}}_{Cs} - \bar{\mathbf{x}}_{Cr}\right)' \mathbf{T}^{-1}_{\varepsilon\varepsilon r}\sum_r d_k y_k\boldsymbol{\varepsilon}_k.\end{aligned}$$

Notice that with perfect balance the explaining power of M-variables is depleted and calibrating on these variables is equivalent to the expansion estimator (2).

4.2. Monitoring weights. With the same setup of $\mathbf{x}_k = \binom{\mathbf{x}_{Mk}}{\mathbf{x}_{Ck}}$, the monitoring weights f_k, defined by (5), can be expressed using partitioned matrices

$$f_k = \begin{pmatrix} \mathbf{T}_{Mr} \\ \mathbf{T}_{Cr} \end{pmatrix}' \begin{pmatrix} \mathbf{T}_{MMs} & \mathbf{T}_{MCs} \\ \mathbf{T}_{CMs} & \mathbf{T}_{CCs} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_{Mk} \\ \mathbf{x}_{Ck} \end{pmatrix},$$

where $\mathbf{T}_{Mr} = \sum_r d_k\mathbf{x}_{Mk}$, $\mathbf{T}_{Cr} = \sum_r d_k\mathbf{x}_{Ck}$ are design weighted M- and C-information auxiliary totals in the response set and

$$\begin{aligned}\mathbf{T}_{MMs} &= \textstyle\sum_s d_k\mathbf{x}_{Mk}\mathbf{x}'_{Mk}; & \mathbf{T}_{MCs} &= \textstyle\sum_s d_k\mathbf{x}_{Mk}\mathbf{x}'_{Ck}; \\ \mathbf{T}_{CCs} &= \textstyle\sum_s d_k\mathbf{x}_{Ck}\mathbf{x}'_{Ck}; & \mathbf{T}_{CMs} &= \mathbf{T}'_{MCs}.\end{aligned}$$

Proposition 3. *Using the same steps and logic as splitting calibration weights g_k we arrive at*

$$f_k = f_{Mk} + h^*_k,$$

where

$$f_{Mk} = \mathbf{T}'_{Mr}\mathbf{T}^{-1}_{MMs}\mathbf{x}_{Mk}, \tag{16}$$

$$h^*_k = \mathbf{T}^{*\prime}_{\varepsilon r}\mathbf{T}^{*-1}_{\varepsilon\varepsilon s}\boldsymbol{\varepsilon}^*_k. \tag{17}$$

Here $\mathbf{T}^*_{\varepsilon\varepsilon s} = \sum_s d_k\boldsymbol{\varepsilon}^*_k\boldsymbol{\varepsilon}^{*\prime}_k$ and $\mathbf{T}^*_{\varepsilon r} = \sum_r d_k\boldsymbol{\varepsilon}^*_k$, where $\boldsymbol{\varepsilon}^*_k = \mathbf{x}_{Ck} - \mathbf{B}'_s\mathbf{x}_{Mk}$ are residuals of linear estimates of $\mathbf{x}_{Ck}$ regressed on $\mathbf{x}_{Mk}$ over the sample and the matrix of coefficients is $\mathbf{B}_s = \mathbf{T}^{-1}_{MMs}\mathbf{T}_{MCs}$.

Remark. Note that the residuals are summed over the response set and in fact $\mathbf{T}^*_{\varepsilon r} = \sum_r d_k\boldsymbol{\varepsilon}^*_k = \sum_{s-r} d_k\boldsymbol{\varepsilon}^*_k$, since $\sum_s d_k\boldsymbol{\varepsilon}^*_k = \mathbf{0}_q$.

Proposition 4. *The h^* weights in a response set r have the following properties:*

$$\begin{aligned}
\bar{h}^*_s &= \frac{\sum_s d_k h^*_k}{\sum_s d_k} = 0;\\
var_s\left(h^*\right) &= \frac{\sum_s d_k \left(h^*_k - \bar{h}^*_s\right)^2}{\sum_s d_k} = P\bar{h}^*_r;\\
cov_s(g_M, h^*) &= \frac{\sum_s d_k \left(g_{Mk} - \bar{g}_{Ms}\right)\left(h^*_k - \bar{h}^*_s\right)}{\sum_s d_k} = 0;\\
cov_s(f_M, h^*) &= \frac{\sum_s d_k \left(f_{Mk} - \bar{f}_{Ms}\right)\left(h^*_k - \bar{h}^*_s\right)}{\sum_s d_k} = 0,
\end{aligned}$$

where $f_{Mk} = \mathbf{T}'_{Mr}\mathbf{T}^{-1}_{MMs}\mathbf{x}_{Mk}$ *and*

$$\bar{g}_{Mr} = \sum_r d_k g_{Mk} / \sum_r d_k, \quad \bar{f}_{Mr} = \sum_r d_k f_{Mk} / \sum_r d_k.$$

In case of perfect balance on the monitoring variables, response propensities f_{Mk} and h^*-weights simplify to $f_{Mk} = P$ for all k, and

$$h^*_k = \hat{N}\left(\bar{\mathbf{x}}_{Cr} - \bar{\mathbf{x}}_{Cs}\right)' \mathbf{T}^{*-1}_{\boldsymbol{\varepsilon\varepsilon} s} \boldsymbol{\varepsilon}^*_k$$

because

$$\bar{\boldsymbol{\varepsilon}}^*_r = \sum_r d_k \boldsymbol{\varepsilon}^*_k / \sum_r d_k = \left(\bar{\mathbf{x}}_{Cr} - \bar{\mathbf{x}}_{Cs}\right).$$

The imbalance measure in this case (and using Proposition 4) takes the form

$$IMB = \mathrm{var}_s(f_M) + \mathrm{var}_s(h^*) = \left(\bar{\mathbf{x}}_{Cr} - \bar{\mathbf{x}}_{Cs}\right)' \Sigma^{*-1}_{\boldsymbol{\varepsilon\varepsilon} s} \left(\bar{\mathbf{x}}_{Cr} - \bar{\mathbf{x}}_{Cs}\right)$$

and imbalance is thereby the balance of C-information weighted by a matrix of residual information from regressing $\mathbf{x}_{Mk}$ on $\mathbf{x}_{Ck}$.

5. Measuring balance with the split auxiliary vector

With the distinction of auxiliary variables by usage, the imbalance measure used in the monitoring phase of a survey is

$$IMB_M = \mathrm{var}_s(f_M).$$

It can be calculated at any step of the survey process. Following Proposition 4 and (14), the effect of added auxiliary variables in the overall imbalance measure can be separated:

$$IMB = \mathrm{var}_s(f) = \mathrm{var}_s(f_M) + \mathrm{var}_s(h^*) = IMB_M + IMB_C.$$

One can see that additional variables in the imbalance measure always increases the IMB value because $\mathrm{var}_s(h^*) \geq 0$. Indication of extra imbalance

from the additional variables gives a ratio

$$Q = \frac{\mathrm{var}_s\left(h^*\right)}{IMB}. \tag{18}$$

The ratio is large, i.e., closer to 1, if the response set is balanced on the monitoring variables, but the M-information does not explain the C-information, and there is extra imbalance in the response with regard to C-information.

Remark. Currently the auxiliary vector is split by usage of the auxiliary variables, but this can be done arbitrarily with other purposes, for example to separate variables that contribute the most to IMB. Similarly, when the auxiliary vector produces mutually exclusive and exhaustive groups, population groups with the highest influence to the overall imbalance can be identified; the ideas have been discussed in [14] and in [11].

Plugging the new form of g-weights back into the the calibration estimator reveals that using both types of information can be divided into two parts:

$$\begin{aligned}\hat{Y}_{CAL} &= \sum_r d_k y_k \left(g_{Mk} + h_k\right) = \sum_r d_k y_k g_{Mk} + \sum_r d_k y_k h_k \\ &= \hat{Y}_{M,CAL} + \sum_r d_k y_k h_k.\end{aligned}$$

The difference in study variable means (6) can now be broken down by the distinction of auxiliary variables

$$\begin{aligned}\bar{y}_r - \bar{y}_s &= \left[\left(\bar{\mathbf{x}}_{Mr} - \bar{\mathbf{x}}_{Ms}\right)' \mathbf{b}_{Mr} - P\mathrm{cov}_r\left(y, h\right)\right] \\ &\quad + \left[\left(\mathbf{b}_{Mr} - \mathbf{b}_{Ms}\right)' \bar{\mathbf{x}}_{Ms} + P\mathrm{cov}_r\left(y, h\right)\right],\end{aligned} \tag{19}$$

since, because of (12), $\sum_r d_k y_k h_k$ can be interpreted as the covariance between the study variable and h-weights:

$$\sum_r d_k y_k h_k = \left(\sum_r d_k\right) \mathrm{cov}_r\left(y, h\right).$$

Since h-weights are formed from residuals, the covariance shows that extra auxiliary information that is highly correlated with the study variable, but not well explained by M-information, is a good candidate for the effective C-information.

The distinction of auxiliary variables by their role in the whole survey process leads to considering the following strategies:

(1) put effort into collecting more auxiliary information and focus on post-weighting correction;
(2) put effort into monitoring response to get a representative set of respondents.

These strategies are studied in the following simulation study.

6. Simulations

The effects of additional auxiliary information was studied in a simulation study. A population was composed on real data from the Estonian Household Survey, with 14 139 households. The database had the following variables for every household (HH) or head of the household (HHH):

- HH net income (study variable);
- Employment status of HHH (0 – not employed, 1 – employed);
- Gender – gender of HHH (0 – male, 1 – female);
- Education level – highest obtained education level of HHH (lower, middle, higher education);
- Number of children in HH;
- HH size, i.e., number of persons belonging to the same HH;
- HH expenditure.

In order to satisfy the requirements (1) and (9), HHH education level was taken as a group vector – a 3-dimensional indicator vector which indicates the education level group where unit k belongs to. The population total of the study variable, $Y = \sum_U y_k$ is known and is used to evaluate simulation results.

6.1. General setup. A SRS sample of size $n = 700$ was taken and from there a response set of 350 respondents (50% response rate) was randomly picked with a given response mechanism, that deliberately produced non-representative response set (details are presented in next section). Gathering the next 100 respondents (raising the final response rate to 64.3%) was done with two different strategies.

Strategy 1: Response set was allowed to accumulate as it had done previously – according to the predetermined response mechanism.

Strategy 2: Response accumulation was monitored and the process intervened by using the fixed proportion method [11].

In the end there were two response sets with the same response rate, one with no intervention done and non-representative, and one where the response set accumulation was monitored and intervened. In the case of Strategy 2 HHH education level, HH size and HHH gender were used for monitoring the response. Details of the monitoring procedure are brought out in a later section.

In the estimation stage four different auxiliary vectors with different combinations of C-information were compared to capture four situations of interest.

	$\mathbf{x}_{Mk}$	$\mathbf{x}_{Ck}$
Vector 1	Education + Gender + HH size	–
Vector 2	Education + Gender + HH size	No. of children
Vector 3	Education + Gender + HH size	HH expenditure
Vector 4	Education + Gender + HH size	No. of children + Empl. status + HH expenditure

In the case of Vector 1 there was no extra C-information, only $\mathbf{x}_{Mk}$ was used in the estimation stage and it serves as a basis for comparisons. For Vector 2 number of children in HH was added to the auxiliary vector in the estimation stage since it is strongly correlated with HH size (correlation 0.72). HH size was used in monitoring stage and was therefore influential in forming $\mathbf{B}_r$ and $\mathbf{B}_s$. For Vector 3 the C-information included HH expenditure since it had a stronger correlation with the study variable (correlation 0.65) than any other auxiliary variable, therefore influential in forming $\mathbf{b}_r$ and $\mathbf{b}_s$. The last vector setup included all of the auxiliary information.

This setup was repeated for 1000 times and for every repetition, strategy, and vector $\hat{Y}_{CAL}$ was calculated with the corresponding auxiliary vector. To assess the performance of estimators relative bias (RB) and relative root mean square error (RRMSE) of the estimated population totals were calculated over all repetitions:

$$\begin{aligned} \mathrm{RB}(\hat{Y}_{CAL}) &= \frac{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{Y}_{CAL}^{t} - Y\right)}{Y}, \\ \mathrm{RRMSE}(\hat{Y}_{CAL}) &= \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{Y}_{CAL}^{t} - Y\right)^{2}}}{Y}, \end{aligned}$$

where T is the number of repetitions, $\hat{Y}_{CAL}^{t}$ is the estimated total of the study variable in t-th repetition, and $Y = \sum_U y_k$ is the true population total.

6.2. Response mechanism. A known response mechanism was generated by computing response probabilities θ_k with the formula

$$logit(\theta_k) = 5 - 4\times\text{HHH gender} + 2\times\text{HHH empl. status} - 0.0004\times\text{HH income}.$$

With this model HH-s with lower income had higher odds of belonging in the response set than HH-s with higher income, female HHH had a lower chance of responding than male HHH, and employed HHH had a higher chance of responding than unemployed HHH. The known response probability for each sampled HH was then multiplied by a random number z_k from a uniform distribution, $z \sim U(0,1)$. The result was sorted in a descending order and top 350 units were taken as the starting point for Strategy 2, and top 450 as the final set of respondents for Strategy 1.

The choice of variables in the response influencing model was done so that it would include one M-information and one C-information variable and the study variable.

6.3. Monitoring and intervening the response accumulation. The fixed proportion method was proposed by [11]. This method sets aside a fixed proportion of the sample at each intervention point. The number of intervention points L needs to be determined in advance, in current simulations there were 10 intervention points. At each intervention point response propensities (16) were computed for all $k \in s$. The values were ordered, and $100/(L+1)$ percent of units with the highest values of f_{Mk} and not belonging to the response set, were set aside, pretending that data collection attempts had stopped for these units. From the remaining set of non-respondents 10 units were randomly assigned to the response set using previously computed response probabilities θ_k and random numbers z_k analogously to the forming of initial response set.

6.4. Results. Vector 1 with Strategy 2 corresponds to the situation where no extra information is acquired, the response is monitored using M-information and calibrated on the same auxiliary variables. Vectors 2–4 with Strategy 1 are situations where additional auxiliary variables are added to the calibration vector, but no prior intervention with the data collection process has been done. Vectors 2–4 with Strategy 2 would be the ideal cases where data collection is monitored and also additional explaining power is acquired.

To assess the effect of additional C-information on the overall imbalance, the means of IMB_C and Q over repetitions was computed, where Q is defined in (18). For variance of the total estimate $\hat{Y}_{CAL}$ over simulations, coefficient of variation was calculated.

Table 1. Simulation results.

	RB		RRMSE		IMB_C		$\bar{Q}$	
	Str 1	Str 2	Str 1	Str 2	Str 1	Str 2	Str 1	Str 2
Vector 1	-.246	-.224	.248	.226	-	-	-	-
Vector 2	-.243	-.222	.245	.223	.001	.001	5.5%	7.5%
Vector 3	-.199	-.175	.201	.177	.012	.012	33.1%	42.4%
Vector 4	-.205	-.182	.208	.184	.015	.014	37.0%	47.2%

Simulation results, presented in Table 1, generated the following comments:

- Results of Propositions 2 and 4 were confirmed, covariances of h-weights with f_{Mk} and g_{Mk} in the response set, and covariances of h^*-weights with f_{Mk} and g_{Mk} in the sample were all 0. Mean of

h-weights in the response set, and mean of h^*-weights in the sample were 0.

- All of the monitored auxiliary vectors produced a lower IMB_M value as expected – for Strategy 1 it was 0.025 and for Strategy 2 around 0.017.
- When comparing Strategy 1 versus Strategy 2, RB and RRMSE were always lower when the same auxiliary vector is used in estimation, meaning that monitoring response on average gave less biased and more accurate estimates.
- Most of the deviance of estimates from the true value was caused by bias since RRMSE values were bigger, but always close to RB values.
- Acquiring strong auxiliary information like HH expenditure gave on average lower bias and lower variation in estimates, more than only monitoring data collection and no additional explaining power in the estimation stage (Vector 3–4 with Strategy 1 versus Vector 1 with Strategy 2 in Table 1).
- Balancing on the M-variables slightly improved the balance of C-variables, as IMB_C decreased when comparing Strategy 1 and 2 through Vector 2–4, but the difference is so small that it cannot be witnessed in Table 1.
- Proportion of imbalance indicated by C-information is higher for auxiliary vectors with Strategy 2, but this is because the overall IMB is lower for Strategy 2.
- Coefficient of variation of $\hat{Y}_{CAL}$ over the simulations did not change significantly when comparing Strategy 1 and Strategy 2 through Vectors 1 to 4, but did increase with adding more auxiliary variables being 2.9% with Vectors 1 and 2, 3.6% with Vectors 3 and 4.

7. Conclusions

The expression for the calibration weights with two terms were derived, explicitly showing the contribution of M-information and C-information to the calibration estimator $\hat{Y}_{CAL}$. The additional explaining power was expressed through h-weights and useful properties, like the mean, variance and covariances of the h-weights were proved.

The expression of response propensities in a split vector case was also developed and some useful properties of h^*-weights were proved. A measure of extra imbalance and its contribution to the overall imbalance was proposed.

A simulation study with a fairly realistic setup, with real data and two alternative strategies was carried out. Simulations confirmed the theoretical results and also showed that monitoring data collection improves estimates in terms of bias and accuracy. If there is strong auxiliary information made available in the estimation stage, then monitoring data collection can be

skipped, although not advised. If the survey budget allows both data collection monitoring and obtaining more auxiliary information, then this would be the most advised course of action, as the simulations demonstrated.

The complementary theoretical results and simulation study are a good starting point for statistical organisations to explore possible strategies for data collection and/or acquisition of extra explaining power. Increased costs force survey statisticians to improve the data collection strategies and measure the data inflow with better indicators. Decisions for survey designs can be drawn from earlier surveys using the ratio of imbalance Q introduced by added explaining power and results of Propositions 1 and 3.

Results of Propositions 2 and 4, and (19) serve as a good mid-point for further development of a theoretical link between reduction of imbalance and non-response bias.

Appendix: Proof of Proposition 1

The proof follows a technique used in [13], where calibration weights are developed into two parts for domain estimation.

When the auxiliary vector is split like in (8), then calibration weights (4) can be expressed using partitioned matrices

$$g_k = \begin{pmatrix} \mathbf{T}_{Ms} \\ \mathbf{T}_{Cs} \end{pmatrix}' \begin{pmatrix} \mathbf{T}_{MMr} & \mathbf{T}_{MCr} \\ \mathbf{T}_{CMr} & \mathbf{T}_{CCr} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_{Mk} \\ \mathbf{x}_{Ck} \end{pmatrix}.$$

Using partitioned matrix inversion for symmetric matrices [4, pp. 74–75], we get

$$\begin{pmatrix} \mathbf{T}_{MMr} & \mathbf{T}_{MCr} \\ \mathbf{T}_{CMr} & \mathbf{T}_{CCr} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{T}_{MMr}^{-1} + \mathbf{B}_r \left[\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r\right]^{-1} \mathbf{B}_r' & -\mathbf{B}_\mathbf{r} \left[\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r\right]^{-1} \\ -\left[\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r\right]^{-1} \mathbf{B}_r' & \left[\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r\right]^{-1} \end{pmatrix},$$

where $\mathbf{B}_r = \mathbf{T}_{MMr}^{-1}\mathbf{T}_{MCr}$ is a coefficient matrix for a multivariate multiple regression fit [7, pp. 279–281]. The calibration weights g_k can now be written in the following way:

$$\begin{aligned} g_k &= \begin{pmatrix} \mathbf{T}_{Ms}'\mathbf{T}_{MMr}^{-1} - \left[\mathbf{T}_{Cs}' - \mathbf{T}_{Ms}'\mathbf{B}_r\right]\left[\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r\right]^{-1}\mathbf{B}_r' \\ \left[\mathbf{T}_{Cs}' - \mathbf{T}_{Ms}'\mathbf{B}_r\right]\left[\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r\right]^{-1} \end{pmatrix}' \begin{pmatrix} \mathbf{x}_{Mk} \\ \mathbf{x}_{Ck} \end{pmatrix} \\ &= \mathbf{T}_{Ms}'\mathbf{T}_{MMr}^{-1}\mathbf{x}_{Mk} \\ &\quad + \left[\mathbf{T}_{Cs}' - \mathbf{T}_{Ms}'\mathbf{B}_r\right]\left[\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r\right]^{-1}\left[\mathbf{x}_{Ck} - \mathbf{B}_r'\mathbf{x}_{Mk}\right]. \end{aligned} \tag{20}$$

Regression coefficients $\mathbf{B}_r$ are used to obtain linear predictions for $\mathbf{x}_{Ck}$ from $\mathbf{x}_{Mk}$, $\hat{\mathbf{x}}_{Ck} = \mathbf{B}_r'\mathbf{x}_{Mk}$, and the middle expressions in brackets can be expressed

via residuals:

$$\begin{aligned}\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r &= \sum_r d_k \mathbf{x}_{Ck}\mathbf{x}'_{Ck} - \sum_r d_k \mathbf{x}_{Ck}\mathbf{x}'_{Mk}\mathbf{B}_r \\ &= \sum_r d_k \mathbf{x}_{Ck}\left(\mathbf{x}_{Ck} - \hat{\mathbf{x}}_{Ck}\right)' \\ &= \sum_r d_k \mathbf{x}_{Ck}\boldsymbol{\varepsilon}'_k, \end{aligned} \tag{21}$$

where $\boldsymbol{\varepsilon}_k = \mathbf{x}_{Ck} - \hat{\mathbf{x}}_{Ck}$ is the vector of residuals. Since $\sum_r d_k \hat{\mathbf{x}}_{Ck}\boldsymbol{\varepsilon}'_k = \sum_r d_k \hat{\mathbf{x}}_{Ck}\left(\mathbf{x}_{Ck} - \hat{\mathbf{x}}_{Ck}\right)' = \mathbf{B}'_r\left(\mathbf{T}_{MCr} - \mathbf{T}_{MMr}\mathbf{B}_r\right) = \mathbf{0}_q$, then an alternative expression for (21) is

$$\mathbf{T}_{CCr} - \mathbf{T}_{CMr}\mathbf{B}_r = \sum_r d_k \mathbf{x}_{Ck}\boldsymbol{\varepsilon}'_k = \sum_r d_k \boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}'_k =: \mathbf{T}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}r}.$$

Similarly

$$\begin{aligned}\mathbf{T}'_{Cs} - \mathbf{T}'_{Ms}\mathbf{B}_r &= \sum_s d_k \mathbf{x}'_{Ck} - \sum_s d_k \mathbf{x}'_{Mk}\mathbf{B}_r \\ &= \sum_s d_k \left(\mathbf{x}_{Ck} - \hat{\mathbf{x}}_{Ck}\right)' \\ &= \sum_s d_k \boldsymbol{\varepsilon}'_k \\ &=: \mathbf{T}'_{\boldsymbol{\varepsilon}s}. \end{aligned}$$

Inserting the matrices $\mathbf{T}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}r}$ and $\mathbf{T}_{\boldsymbol{\varepsilon}s}$ into (20) we get

$$g_k = \mathbf{T}'_{Ms}\mathbf{T}^{-1}_{MMr}\mathbf{x}_{Mk} + \mathbf{T}'_{\boldsymbol{\varepsilon}s}\mathbf{T}^{-1}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}r}\boldsymbol{\varepsilon}_k$$

and taking $g_{Mk} := \mathbf{T}'_{Ms}\mathbf{T}^{-1}_{MMr}\mathbf{x}_{Mk}$, $h_k := \mathbf{T}'_{\boldsymbol{\varepsilon}s}\mathbf{T}^{-1}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}r}\boldsymbol{\varepsilon}_k$ we arrive at the result of Proposition 1.

Acknowledgement

The author would like to thank Associate Professor Imbi Traat for her many suggestions, Professor Carl-Erik Särndal for consultations on the matter, and the anonymous referee for the constructive comments and suggestions. This work was supported by institutional research funding IUT34-5 of the Estonian Ministry of Education and Research.

References

[1] R. R. Andridge and R. J. A. Little, *Proxy pattern-mixture analysis for survey nonresponse*, J. Off. Stat. **27** (2011), 153–180.

[2] J.-C. Deville and Y. Tillé, *Efficient balanced sampling: the cube method*, Biometrika **91**(4) (2004), 893–912.

[3] R. M. Groves and S. G. Heeringa, *Responsive design for household surveys: tools for actively controlling survey errors and costs*, J. R. Statist. Soc. A **169**(3) (2006), 439–457.
[4] T. Kollo and D. von Rosen, *Advanced Multivariate Statistics with Matrices*, Springer, Dordrecht, 2005.
[5] F. Kreuter, *Improving Surveys with Paradata: Analytic Uses of Process Information*, John Wiley & Sons, Inc., Hoboken, 2013.
[6] P. Lundquist and C.-E. Särndal, *Aspects of responsive design with applications to the Swedish living conditions survey*, J. Off. Stat. **29** (2013), 557–582.
[7] A. C. Rencher, *Multivariate Statistical Inference and Application*, Wiley, New York, 1998.
[8] C.-E. Särndal, *The 2010 Morris Hansen lecture dealing with survey nonresponse in data collection, in estimation*, J. Off. Stat. **27** (2011), 1–21.
[9] C.-E. Särndal, K. Lumiste, and I. Traat, *Reducing the response imbalance: Is the accuracy of the survey estimates improved?*, Surv. Methodol. **42** (2016), 219–238.
[10] C.-E. Särndal and P. Lundquist, *Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation*, J Surv. Stat. Methodol. **2** (2014), 361–387.
[11] C.-E. Särndal and P. Lundquist, *Balancing the response and adjusting estimates for nonresponse bias: complementary activities*, J. SFdS **155**(4) (2014), 28–50.
[12] C.-E. Särndal and S. Lundström, *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Ltd., Chichester, 2005.
[13] C.-E. Särndal and I. Traat, *Domain estimators calibrated on information from another survey*, Acta Comment. Univ. Tartu. Math. **15** (2011), 43–60.
[14] B. Schouten, J. Bethlehem, K. Beullens, Ø. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner, *Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators*, Int. Stat. Rev. **80** (2012), 382–399.
[15] B. Schouten, F. Cobben, and J. Bethlehem, *Indicators for the representativeness of survey response*, Surv. Methodol. **35** (2009), 101–113.
[16] R. Tourangeau, J. M. Brick, S. Lohr, and J. Li, *Adaptive and responsive survey designs: a review and assessment*, J. R. Statist. Soc. A **180(1)** (2017), 203–223.

Institute of Mathematics and Statistics, University of Tartu, J. Liivi 2-517, 50409 Tartu, Estonia

E-mail address: kaur.lumiste@ut.ee