METRON (2026) 84:29–43
https://doi.org/10.1007/s40300-025-00303-z

A Hybrid GREG Estimator for Estimation in Cluster Sampling

Anne Konrad[1] · **Jan Pablo Burgard**[2]

Received: 3 July 2023 / Accepted: 10 December 2025 / Published online: 22 December 2025

Abstract
Cluster sampling is a widely used sampling design for survey data selection. The collected data provide information on both the individual units and the clusters. The parameter of interest is an unknown population total at the unit level. Due to the cluster sampling design, the model-assisted generalized regression estimator can be established either at the unit or at the cluster level. The unit-level estimator, however, ignores the cluster sampling design. The cluster-level estimator relies solely on the per-cluster aggregated information, which leads to the loss of individual patterns and the risk of ecological fallacy. As a remedy, in this paper we propose a hybrid generalized regression estimator as a new approach that balances between unit- and cluster-level modeling. It is implemented at the unit level like the study variable. However, in addition to the information of the units, it also takes into account the information of the other cluster members.

Keywords Generalized regression estimator · Cluster sampling · Design-based inference

1 Introduction

Cluster sampling is a widely used sampling design where the finite population is divided into subpopulations of units, called clusters. Potential subpopulations may be geographical districts, city blocks, enterprises, establishments, schools, or some other type of aggregate unit. Cluster sampling is often implemented when no complete, up-to-date and accessible list of all population units is available to serve as a sampling frame [8, p.243]. A further reason to implement cluster sampling is that the costs of selecting clusters might be negligible compared to the costs of selecting the individual units [17, p.170]. For instance, dividing the population into clusters can reduce the travel costs of interviewers [28, p.203]. In the sampling process, the clusters are sampled in a first stage. In a second stage, the units are selected from the sampled clusters. Therefore, the sample information is provided both at the unit and the cluster level.

✉ Anne Konrad
anne.konrad@lifbi.de

Jan Pablo Burgard
burgardj@uni-trier.de

[1] Statistical Survey Methods, Leibniz Institute for Educational Trajectories, Wilhelmplatz 3, Bamberg 96047, Germany

[2] Economic and Social Statistics, Trier University, Universitaetsring 15, Trier 54296, Germany

The objective to estimate is an unknown population total at the unit level. A well-known model-assisted estimator is the generalized regression (GREG) estimator, which incorporates auxiliary information in the estimation process. Under cluster sampling and when estimating unit-level characteristics, the assisting model of the GREG estimator can either be established at the unit or the cluster level. The unit-level GREG estimator captures only the information of the auxiliary variables and the study variable of the units themselves. Thus, the point estimator does not consider the cluster sampling design. The cluster-level GREG estimator, in turn, considers the per-cluster aggregated unit-level information of the auxiliaries and the study variable. However, through the aggregation, individual patterns are lost, which is particularly problematic when the clusters tend to be heterogeneous. Moreover, the aggregation per cluster can lead to ecological fallacy when the relationship between the study and auxiliary variables differs between both levels. Ecological fallacy occurs when aggregation per cluster leads to incorrect conclusions about the true relationship between the variables, resulting in an incorrect statistical inference and a loss of efficiency of the estimator.

To balance between the issues of unit-level and cluster-level modeling, we propose a hybrid GREG estimator. The proposed estimator is implemented at the unit level and thus utilizes individual study and auxiliary variables rather than aggregated information. Additionally, it incorporates information from the other cluster members over a convex combination. The extent to which the cluster members' information is used is determined by weighting factors. Therefore, the proposed hybrid GREG estimator integrates the cluster sampling design within the point estimation process, while avoiding aggregation of unit-level information.

The remainder of the paper is organized as follows: Section 2 introduces the unit-level GREG estimator. Section 3 discusses cluster-level GREG estimators. In Section 4, we derive the proposed hybrid GREG estimator and its properties. A Monte Carlo simulation study compares the performance of the hybrid, the unit- and the cluster-level GREG estimators (Section 5). Section 6 summarizes the results.

2 Unit-Level GREG Estimator Under Cluster Sampling

In cluster sampling, the finite population is divided into disjoint subpopulations, so-called clusters. It is distinguished into single-stage and multi-stage cluster sampling. In single-stage cluster sampling, a sample of clusters is selected from the finite population and all units within the selected clusters are surveyed. Multi-stage cluster sampling consists of two or more stages of probability sampling [27]. For simplicity, we assume single-stage cluster sampling and use the terms single-stage cluster sampling and cluster sampling synonymously.

In single-stage cluster sampling, a sample s_I is selected from the finite population of clusters $U_I = \{1, \ldots, g, \ldots, M\}$ according to the sampling design $p(\cdot)$, where $p(s_I)$ is the probability of selecting s_I. The sample size of s_I is m. Let U_g be the population of units within a cluster g of size N_g. The sampling design generates for every cluster g a known first-order inclusion probability $\pi_g = Pr(g \in s_I)$ with $\pi_g > 0$. The second-order inclusion probability of cluster g and k is denoted as $\pi_{gk} = Pr(g, k \in s_I)$. Note that $\pi_{gg} = \pi_g$. The finite population of units is denoted by $U = \{1, \ldots, i, \ldots, N\}$. The corresponding sample of the units is $s = \cup_{g \in s_I} U_g$ with sample size n. Each unit in the selected clusters is sampled, which implies that $\pi_i = Pr(i \in s) = Pr(g \in s_I) = \pi_g$.

The objective is to estimate the unknown population total of a unit-level study variable y defined by

$$\tau_y = \sum_{i \in U} y_i, \tag{1}$$

where y_i is the value of y for unit i. A design-unbiased estimator for τ_y is the Horvitz-Thompson (HT) estimator [10, 22]. If the sample is selected according to a cluster sampling design, the HT estimator is given by

$$\widehat{\tau}_y = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{g \in s_I} \frac{y_g}{\pi_g} \tag{2}$$

with $y_g = \sum_{i \in U_g} y_i$. It is valid that $\sum_{i \in s} y_i = \sum_{g \in s_I} \sum_{i \in U_g} y_i$.

The efficiency of the estimates can be improved by the use of auxiliary information in the estimation process. A widely used model-assisted estimator incorporating auxiliary information is the GREG estimator established by [8], [4], [26], [13] and [29]. The auxiliary vector of unit i is given by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq}, \ldots, x_{iQ})^\top$ where $x_{i1} = 1$ determines the intercept. The corresponding total vector $\boldsymbol{\tau}_x = \sum_U \boldsymbol{x}_i$ of dimension Q is assumed to be known from censuses, registers, or other reliable sources.

The GREG estimator relies on a linear regression model ξ given by

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i \quad \text{for all} \quad i \in U \tag{3}$$

that specifies the relationship between the study variable and the auxiliary variables. Model (3) is defined at the unit level. Here, $\boldsymbol{\beta}$ is a superpopulation regression coefficient, and ϵ_i is an unobserved random error. Note that $E_\xi(\epsilon_i) = 0$, $V_\xi(\epsilon_i) = v_i \sigma^2$ and $\mathrm{E}_\xi(\epsilon_i \epsilon_j) = 0$ for all $i \neq j$. E_ξ and V_ξ denote the expectation and the variance with respect to the model ξ. The variance parameter v_i, with $v_i > 0$, has to be known and can be used to capture heteroscedasticity. In order to simplify notation, we assume $v_i = 1$ (homoscedasticity).

Define $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ as $(n \times Q)$-matrix, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ as $(n \times 1)$-vector and $\boldsymbol{\Pi} = \mathrm{diag}(\pi_1, \ldots, \pi_n)$ as $(n \times n)$-matrix. Then, the linear unit-level GREG estimator under cluster sampling relying on (3) is defined by

$$\widehat{\tau}_y^{\mathrm{UNIT}} = \widehat{\tau}_y + (\boldsymbol{\tau}_x - \widehat{\boldsymbol{\tau}}_x)^\top \widehat{\boldsymbol{B}}^{\mathrm{UNIT}}, \tag{4}$$

where

$$\widehat{\boldsymbol{B}}^{\mathrm{UNIT}} = (\boldsymbol{X}^\top \boldsymbol{\Pi}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Pi}^{-1} \boldsymbol{y} \tag{5}$$

is a weighted least squares estimate for β [27, pp.307]. Based on the Taylor linearization, [26] showed that the GREG estimator is design-consistent. [5] proved the approximately design-unbiasedness under mild design conditions for the assisting model and for the sampling design.

Define $\boldsymbol{L}$ as a $(n \times m)$-matrix with entity l_{ig} has value one if unit i belongs to cluster g and zero otherwise. Then, summing up the unit-level quantities per cluster results in

$$\boldsymbol{X}_c = \boldsymbol{L}^\top \boldsymbol{X} \quad \text{and} \quad \boldsymbol{y}_c = \boldsymbol{L}^\top \boldsymbol{y}.$$

The variance estimator of (4) under cluster sampling approximated by Taylor linearization can be obtained from

$$\widehat{V}(\widehat{\tau}_y^{\mathrm{UNIT}}) = (\boldsymbol{y}_c - \boldsymbol{X}_c^\top \widehat{\boldsymbol{B}}^{\mathrm{UNIT}})^\top \boldsymbol{\Pi}_c^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi}_c^{-1} (\boldsymbol{y}_c - \boldsymbol{X}_c^\top \widehat{\boldsymbol{B}}^{\mathrm{UNIT}}) \tag{6}$$

with $\boldsymbol{\Pi}_c = \text{diag}(\pi_1, \ldots, \pi_m)$ and $\boldsymbol{\Delta}_c = \{\frac{\pi_{gk} - \pi_g \pi_k}{\pi_{gk}}\}$ [27, p.129, p.235]. It becomes obvious that in the case of cluster sampling the variance of the unit-level GREG estimator depends on $\boldsymbol{y}_c$ and $\boldsymbol{X}_c$. That is true, although the assisting model ξ, defined in (3), refers to their unit-level counterparts $\boldsymbol{y}$ and $\boldsymbol{X}$. The point estimator, (4), in contrast, is treated as if the sample was drawn by a unit-level sampling design. Therefore, the cluster sampling design is only considered in the variance estimation, whereas the unit-level point estimator remains unaffected by the cluster sampling design.

3 Cluster-Level GREG Estimators

Under cluster sampling, the unknown unit-level total τ_y, defined in (1), can alternatively be estimated at the cluster level. For a fair comparison with the unit-level GREG estimator, we assume that the same set of auxiliary variables is used at both levels. This is not a restrictive assumption, as variables that are only available at one level can easily be transferred to the respective other level by aggregating the unit-level variables within clusters or by assigning the cluster mean value to each cluster member.

The assisting model at cluster-level is given by

$$y_g = \boldsymbol{x}_g^\top \boldsymbol{\beta} + \epsilon_g \quad \text{for all} \quad g \in U_I \tag{7}$$

with $E_\xi(\epsilon) = 0$, $V_\xi(\epsilon) = v_g \sigma^2$ with $v_g = \sum_{i \in U_g} v_i$, and $\text{E}_\xi(\epsilon_g \epsilon_k) = 0$ for all $g \neq k$. In order to maintain a fair comparison between the unit- and cluster-level estimators, the assumption of homoscedasticity ($v_i = 1$) in the unit level model (3) implies that $v_g = N_g$ at the cluster level. This results in a heteroscedastic model when cluster sizes vary. Then, the cluster-level GREG estimator is obtained from

$$\widehat{\tau}_y^{\text{CLU}} = \widehat{\tau}_y + (\boldsymbol{\tau}_x - \widehat{\boldsymbol{\tau}}_x)^\top \widehat{\boldsymbol{B}}^{\text{CLU}}, \tag{8}$$

where

$$\widehat{\boldsymbol{B}}^{\text{CLU}} = (\boldsymbol{X}_c^\top \boldsymbol{\Pi}_c^{-1} \boldsymbol{V}_c^{-1} \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^\top \boldsymbol{\Pi}_c^{-1} \boldsymbol{V}_c^{-1} \boldsymbol{y}_c \tag{9}$$

with $\boldsymbol{V}_c = \text{diag}(N_1, \ldots, N_g, \ldots, N_m)$. Note that instead of an intercept, $\boldsymbol{X}_c$ contains the number of the units within a cluster, N_g. The unit-level GREG estimator (4) differs from (8) only in terms of the regression coefficient, which is based on the variables aggregated per cluster and not on the unit-level variables. The variance formula of the cluster-level GREG estimator is given by

$$\widehat{V}(\widehat{\tau}_y^{\text{CLU}}) = (\boldsymbol{y}_c - \boldsymbol{X}_c^\top \widehat{\boldsymbol{B}}^{\text{CLU}})^\top \boldsymbol{\Pi}_c^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi}_c^{-1} (\boldsymbol{y}_c - \boldsymbol{X}_c^\top \widehat{\boldsymbol{B}}^{\text{CLU}}).$$

An alternative to (8) is the Montanari GREG estimator. It has minimum variance in the class of linear estimators. The Montanari estimator is obtained by setting the first derivative of the variance of the unit-level GREG estimator (6) with respect to the regression coefficient to zero, which yields the optimal coefficient

$$\boldsymbol{B}^{\text{MON}} = \text{V}(\widehat{\boldsymbol{\tau}}_x)^{-1} \text{Cov}(\widehat{\boldsymbol{\tau}}_x, \widehat{\tau}_y) \tag{10}$$

with Cov as covariance [18]. (10) is optimal in the sense of minimizing the asymptotic variance of the unit-level GREG estimator for large sample sizes [19]. By substituting the variance and covariance by their HT estimates, we obtain

$$\widehat{\boldsymbol{B}}^{\text{MON}} = (\boldsymbol{X}_c^\top \boldsymbol{\Pi}_c^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi}_c^{-1} \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^\top \boldsymbol{\Pi}_c^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi}_c^{-1} \boldsymbol{y}_c. \tag{11}$$

The sampling design is accounted for through first- and second-order inclusion probabilities in $\boldsymbol{\Delta}_c$ [21]. Because $\widehat{\boldsymbol{B}}^{\text{MON}}$ is estimated, it is only approximately optimal [7, p. 3].

The Montanari estimator is given by

$$\widehat{\tau}_y^{\text{MON}} = \widehat{\tau}_y + (\boldsymbol{\tau}_x - \widehat{\boldsymbol{\tau}}_x)^\top \widehat{\boldsymbol{B}}^{\text{MON}}. \tag{12}$$

Consequently, minimizing the variance of the unit-level GREG estimator yields an optimal estimator that refers to the per-cluster aggregates $\boldsymbol{X}_c$ and $\boldsymbol{y}_c$ instead of $\boldsymbol{X}$ and $\boldsymbol{y}$ as used in the underlying unit-level model (3) since the variance of the unit-level GREG estimator under cluster sampling (6) already depends on the cluster aggregates.

[20] showed that the Montanari estimator implicitly relies on an assisting model that includes, in addition to the auxiliary variables, so-called design-balanced variables. A design-balanced variable is any non-zero auxiliary variable that is proportional to the first-order inclusion probabilities within subpopulations and whose mean is estimated without error by the HT estimator [21]. For example, in a stratified random sampling design, design-balanced variables correspond to indicator variables for stratum membership. These additional design-balanced variables may cause instabilities in the Montanari estimator for finite sample sizes [20].

The variance formula of the Montanari estimator is given by

$$\widehat{V}(\widehat{\tau}_y^{\text{MON}}) = (\boldsymbol{y}_c - \boldsymbol{X}_c^\top \widehat{\boldsymbol{B}}^{\text{MON}})^\top \boldsymbol{\Pi}_c^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi}_c^{-1} (\boldsymbol{y}_c - \boldsymbol{X}_c^\top \widehat{\boldsymbol{B}}^{\text{MON}}).$$

The aggregation of the unit-level variables within clusters considers information given by all other cluster members, however, individual patterns are lost. This can be particularly problematic when clusters are heterogeneous. For example, if the clusters are areas comprising both social buildings and detached houses, and the study variable is volatile, such as income, the resulting estimates might be inaccurate if only the aggregated cluster information is used. This issue could be exacerbated when the clusters become larger.

A further issue is that the correlation concerning the same variables can differ at the individual and the aggregated level [25]. Therefore, aggregating the unit-level information per cluster can lead to wrong conclusions about the true relationship between the auxiliaries and the study variable and results in an incorrect regression coefficient. The wrong inference is known as ecological fallacy. As a consequence, the efficiency is decreased, because even though the GREG estimator is asymptotically design-unbiased irrespective of the correctness of the assisting model, its efficiency depends on the explanatory power of the model [27, p. 227, p. 239].

A practical problem of implementing the Montanari estimator is that the second-order inclusion probabilities are often unknown or complex, as in the case of multistage designs [9]. Consequently, (12) is rarely used in practice [2].

4 Proposed Hybrid GREG Estimator

To balance between the above discussed issues of unit- and cluster-level modeling, we propose a hybrid GREG estimator. It is implemented at the unit level which avoids ecological fallacy by ensuring both the objective to be estimated and the assisting model are on the same level.

We define the proposed hybrid GREG estimator as

$$\widehat{\tau}_y^{\text{HYB}} = \widehat{\tau}_y + (\boldsymbol{\tau}_x - \widehat{\boldsymbol{\tau}}_x)^\top \widehat{\boldsymbol{B}}^{\text{HYB}}, \tag{13}$$

where

$$\widehat{\boldsymbol{B}}^{\mathrm{HYB}} = (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{y} \tag{14}$$

is the regression coefficient with $\boldsymbol{W}$ as $(n \times n)$-matrix with weighting elements

$$w_{i,j}^{\mathrm{HYB}} = \begin{cases} \alpha_g \pi_g^{-1} & \text{for } i, j \in U_g \wedge i = j \\ (1-\alpha_g)\pi_g^{-1} & \text{for } i, j \in U_g \wedge i \neq j \\ 0 & \text{for } i \in U_g, j \in U_k, g \neq k \end{cases} \tag{15}$$

with $\alpha_g \in [0.5,\ 1]$. The block matrix structure of $\boldsymbol{W}$ enables the simultaneous use of information from both the auxiliary and the study variables at unit and cluster levels. The extent to which unit and cluster member information is incorporated is determined by α_g. A higher α_g assigns more weight to the unit itself, whereas a smaller α_g enables greater information borrowing from other units within the same cluster. The lower bound of α prevents extreme or unstable weights that could otherwise result in a singular matrix $\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X}$. Moreover, it ensures that the information utilized from the cluster members does not exceed that from the unit itself.

For an auxiliary variable x_{iq}, the information of the cluster members in $(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})$, given by

$$\sum_{g \in s_I} (1-\alpha_g) \sum_{i \in U_g} \sum_{\substack{j \in U_g \\ j \neq i}} \frac{x_{iq}}{\pi_i} \frac{x_{jq}}{\pi_j},$$

can be interpreted as the sum of correlations with respect to pairs of cluster members which we define as *intra-cluster correlations*, assuming that the auxiliaries are standardized. Consequently, the hybrid estimator borrows information from other cluster members in terms of intra-cluster correlations and thus directly incorporates the cluster sampling design into the point estimation.

Analogous to (6), the variance of the proposed hybrid GREG estimator is estimated by

$$\widehat{V}(\widehat{\tau}_y^{\mathrm{HYB}}) = (\boldsymbol{y}_c - \boldsymbol{X}_c^\top \widehat{\boldsymbol{B}}^{\mathrm{HYB}})^\top \boldsymbol{\Pi}^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi}^{-1} (\boldsymbol{y}_c - \boldsymbol{X}_c^\top \widehat{\boldsymbol{B}}^{\mathrm{HYB}}). \tag{16}$$

4.1 Comparison with the Unit- and Cluster-Level Estimators

The unit-level estimator (4), the cluster-level estimators (8) and (12) as well as the hybrid estimator (13) differ only in their regression coefficients, which therefore are the basis of their comparison. When comparing the unit-level coefficient (5) with the hybrid coefficient (14), it becomes obvious that (5) only considers the information of the units themselves and any dependencies among cluster members are ignored. For $\alpha_g = 1$ the other cluster members in (15) are weighted by zero. Consequently,

$$\widehat{\boldsymbol{B}}^{\mathrm{HYB}} = \widehat{\boldsymbol{B}}^{\mathrm{UNIT}}.$$

For a comparison with the cluster-level and the Montanari GREG estimator, we rewrite them at the unit level. (9) can be rewritten as

$$\widehat{\boldsymbol{B}}^{\mathrm{CLU}} = (\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{\Pi}_c^{-1}\boldsymbol{V}_c^{-1}\boldsymbol{L}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{\Pi}_c^{-1}\boldsymbol{V}_c^{-1}\boldsymbol{L}^\top \boldsymbol{y}$$

with $\boldsymbol{W}^{\mathrm{CLU}} = \boldsymbol{L}\boldsymbol{\Pi}_c^{-1}\boldsymbol{V}_c^{-1}\boldsymbol{L}^\top$ as $(n \times n)$-block matrix with weighting elements

$$w_{i,j}^{\mathrm{CLU}} = \begin{cases} \pi_g^{-1} N_g^{-1} & \text{for } i, j \in U_g \\ 0 & \text{for } i \in U_g, j \in U_k, g \neq k. \end{cases}$$

Accordingly, both the hybrid and the cluster-level coefficient capture information from the individual units as well as from the other cluster members. However, only the hybrid estimator allows a different weighting of unit- and cluster-level information. The cluster-level estimator, in turn, assigns the same weight to each type of information. Moreover, the cluster-level estimator assumes heteroscedasticity, whereas the unit-level and hybrid estimators are based on the homoscedasticity assumption.

At the unit level the Montanari coefficient (11) can be expressed as

$$\widehat{\boldsymbol{B}}^{\text{MON}} = (\boldsymbol{X}^\top \boldsymbol{L} \boldsymbol{\Pi}_c^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi}_c^{-1} \boldsymbol{L}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{L} \boldsymbol{\Pi}_c^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi}_c^{-1} \boldsymbol{L}^\top \boldsymbol{y}$$

with $\boldsymbol{W}^{\text{MON}} = \boldsymbol{L} \boldsymbol{\Pi_c}^{-1} \boldsymbol{\Delta}_c \boldsymbol{\Pi_c}^{-1} \boldsymbol{L}^\top$ as fully populated $(n \times n)$-matrix with weighting elements

$$w_{i,j}^{\text{MON}} = \begin{cases} (1-\pi_g)\pi_g^{-2} & \text{for } i, j \in U_g \\ (\pi_{gk} - \pi_g \pi_k)(\pi_g \pi_k \pi_{gk})^{-1} & \text{for } i \in U_g, j \in U_k, g \neq k. \end{cases}$$

Thus, intra-cluster correlations are utilized, as in the hybrid and the cluster-level estimators, which ensures that the sampling design is taken into account. In addition, the Montanari coefficient also incorporates the correlations with respect to pairs of clusters. We name this correlation *cross-cluster correlations*. However, it is unclear to what extent these cross-cluster correlations improve unit-level estimates.

4.2 Choice of the Weighting Factors

Different choices of the weighting factors are possible. As discussed in Section 3, aggregating unit-level variables in the cluster-level estimators can be particularly problematic when the clusters are heterogeneous or large. Accordingly, the weighting factors α_g can be chosen proportional to the heterogeneity or the size of the clusters.

A measure for the heterogeneity is the pooled within-cluster variance across all sampled clusters, obtained from

$$V^{\text{within}}(y) = \frac{\sum_{g \in s_I} (N_g - 1) S_{yU_g}^2}{\sum_{g \in s_I} (N_g - 1)}, \tag{17}$$

where $S_{yU_g}^2 = (N_g - 1)^{-1} \sum_{i \in U_g} (y_i - \bar{y}_g)^2$ is the variance for cluster g and $\bar{y}_g = N_g^{-1} \sum_{i \in U_g} y_i$ is the mean value of cluster g [27, p. 130].

Then, we define

$$\alpha_g = \begin{cases} 1 & \text{for } g \in \{g \in s_I : N_g = 1\} \\ 1 - 0.5 \cdot \frac{S_{yU_g}^2}{S_{yU_g}^2 + V^{\text{within}}(y)} & \text{otherwise.} \end{cases} \tag{18}$$

If $S_{yU_g}^2 = 0$, α_g is set to 1. According to (18), α_g reflects the degree of homogeneity within cluster g relative to the overall within-cluster variation in the sample. Clusters with low within-cluster variance $S_{yU_g}^2$ receive α_g values close to 1, indicating that the hybrid estimator relies primarily on the information of the unit itself. For more heterogeneous clusters, $S_{yU_g}^2$ is relatively large compared to the pooled within-cluster variance $V^{\text{within}}(y)$, resulting in smaller α_g values and allowing the estimator to borrow more information from other cluster members. The definition in (18) ensures that α_g is bounded below by 0.5, such that $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ remains non-singular and even highly heterogeneous clusters retain a minimum contribution

from the unit itself. For clusters with one single unit, there is no information from other cluster members, and thus $\alpha_g = 1$ per definition.

Alternatively, α_g can be chosen according the cluster size, that is

$$\alpha_g = \begin{cases} 1 & \text{for } g \in \{g \in s_I : N_g = 1\} \\ \left(1 - \frac{1}{N_g}\right) & \text{otherwise.} \end{cases} \tag{19}$$

The intention behind this choice is that the larger the cluster, the smaller $(1 - \alpha_g)$, and consequently the lower the weight assigned to the information of all other cluster members. This specification ensures that $\alpha_g \geq 0.5$, thereby maintaining invertibility of $\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X}$. (19) has the advantage of being independent of the study variable y, in contrast to (18). Therefore, the hybrid estimator (13) can be expressed in a linearly weighted form

$$\begin{aligned} \widehat{\tau}_y^{\text{HYB}} &= \boldsymbol{y}^\top [\boldsymbol{\Pi}^{-1}\mathbf{1} + \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{\tau}_x - \widehat{\boldsymbol{\tau}}_x)] \\ &= \boldsymbol{y}^\top \boldsymbol{G} \end{aligned}$$

with $\mathbf{1}$ as n-vector of 1s. The weights $\boldsymbol{G}$, with a_g independent from the study variable y as in (19), are multi-purpose weights because once the auxiliary variables are chosen, it can be applied to any study variable in the survey. This is an essential property for example for official statistics.

5 Simulation Study

In a Monte Carlo (MC) simulation study, we compare the performance of the proposed hybrid, the unit-level, the Montanari and the cluster-level GREG estimator. The simulation study is based on the synthetic and openly accessible data set AMELIA, which is derived from EU-SILC [3]. To reduce the computational burden, we use the data of only one out of four regions. The population consists of approximately 2.6 million persons and 0.9 million households.

In order to study the influence of the heterogeneity of the clusters on the performance of the estimators, we construct heterogeneous and homogeneous clusters (here households) with respect to income from the units (here persons). For the heterogeneous scenario, households are compiled according to the original household identifier (ID) given in the AMELIA data set. Figure 1 illustrates the median income for the different clusters in the population. We plot the median since it is robust against outliers. The upper left plot shows that the median income of the households is very volatile and that the households are heterogeneous. To generate homogeneous households, we redistribute the persons to new households. For this purpose, we sort the persons by income and generate a new household ID by randomly allocating the known distribution of the household sizes from the original household ID to the sorted persons. Thus, the distributions of the household sizes are the same for the heterogeneous and homogeneous scenarios. The upper right plot in Figure 1 shows that the median income is similar for most households, except for those with the highest income.

To examine the influence of the cluster size on the performance of the estimators, we collapse the households to larger clusters. For this purpose, we use the address size distribution of Berlin as a benchmark to generate new address IDs. The procedure is as follows. First, the addresses are sorted by size. Second, households of sizes 1 to 4 are matched to addresses of the same size. Third, for each address with a size greater than 4, we join the first emerging households to addresses such that the sum of the household sizes does not exceed the address

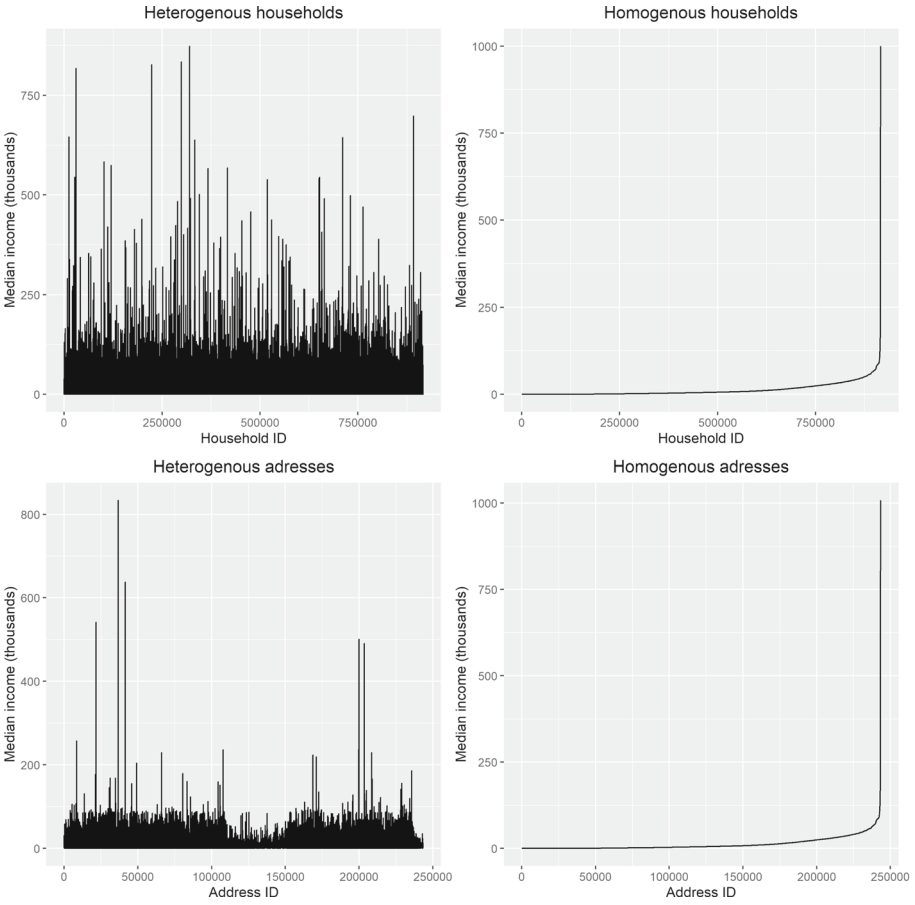

Fig. 1 Median income of the clusters in the population

size. As there are more persons in the AMELIA data set than the sum of address sizes, further steps are necessary. Fourth, we randomly select some address sizes and perform step 3. Fifth, repeat step 4 until all households are allocated to addresses. We denote this scenario as heterogeneous addresses. The lower left plot in Figure 1 confirms the heterogeneity of the addresses with regard to the median income. Analogously to the proceeding in the homogeneous household scenario, we sort the units by income and generate new address IDs by randomly allocating the known distribution of the address sizes from the collapsed original address IDs to the sorted clusters. We call this scenario homogeneous addresses. Once more, the resulting cluster size distributions are the same in both address scenarios. The median income of the homogeneous address clusters is illustrated in the lower right of Figure 1.

Figure 2 plots the distribution of the cluster sizes of the households and addresses in the population. In the right plot, there is a cut-off point at 50 addresses, after which the addresses are grouped into blocks of 100. Because of the grouping, there is a high bar at (50, 150]. For households and addresses, the most common cluster size is 2. The largest household size is 14. The largest address size, in turn, is 1076. As desired, the tails of the address size distribution are much longer than those of the household size distribution. Since the distribution of the

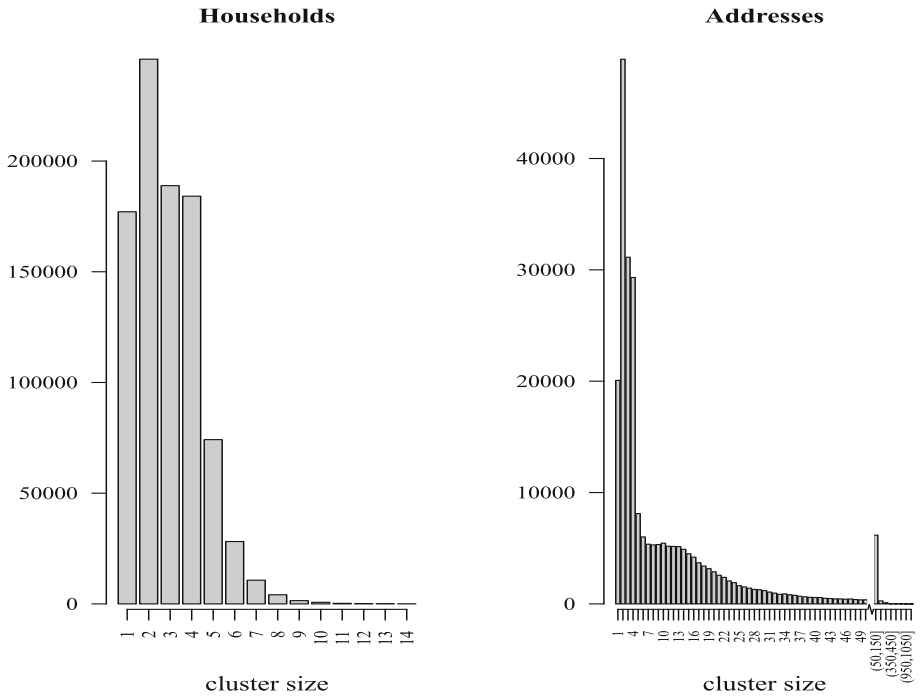

Fig. 2 Barplots of the cluster size of the households and addresses in the population

cluster sizes is the same in the homogeneous and heterogeneous scenarios, it is not necessary to distinguish between the two.

We draw $R = 10,000$ MC samples of different sample sizes using simple random sampling of clusters. In the small sample size scenarios, we select $m = 300$ households and $m = 75$ addresses, resulting in similar sample sizes of units across the settings. The average sample size of the units in the household scenarios is $\bar{n} = 867$, while in the address scenarios, it is $\bar{n} = 817$. In the larger sample size scenario, we select $m = 4.000$ households yielding $\bar{n} = 11.560$ units, and $m = 1.000$ addresses resulting in $\bar{n} = 11.882$ units.

The assisting model of the GREG estimators consists of 4 auxiliary variables, namely: intercept, gender (two categories), age classes (4 categories), and marital status (4 categories). We examine study variables of different types and scales. First, we choose income `inc` and self-employment `sel` as classical unit-level characteristics. Second, we estimate the cross-classification between income and marital status (`inc_ms`), where marital status is an auxiliary variable in the assisting model. In practice, estimates for subgroups or domains are often of as much interest as population totals. Thus, third, we analyze two cross-classifications of age by sex by marital status, which are both of small sizes: `age2_sex1_ms2` (age class 20-39 years, female, married) and `age4_sex0_ms2` (age class 60 years and older, male, married).

For each scenario, we compare the unit-level GREG estimator (UNIT) defined in (4), the cluster-level GREG estimator (CLU) defined in (8), the Montanari GREG estimator (MON) defined in (12) and the hybrid GREG estimator, (13), with weighting factors (18) (HYB_within) and with weighting factors (19) (HYB_Ng). HYB_within is determined by the heterogeneity of the cluster measured by the within variance, and HYB_Ng by the cluster

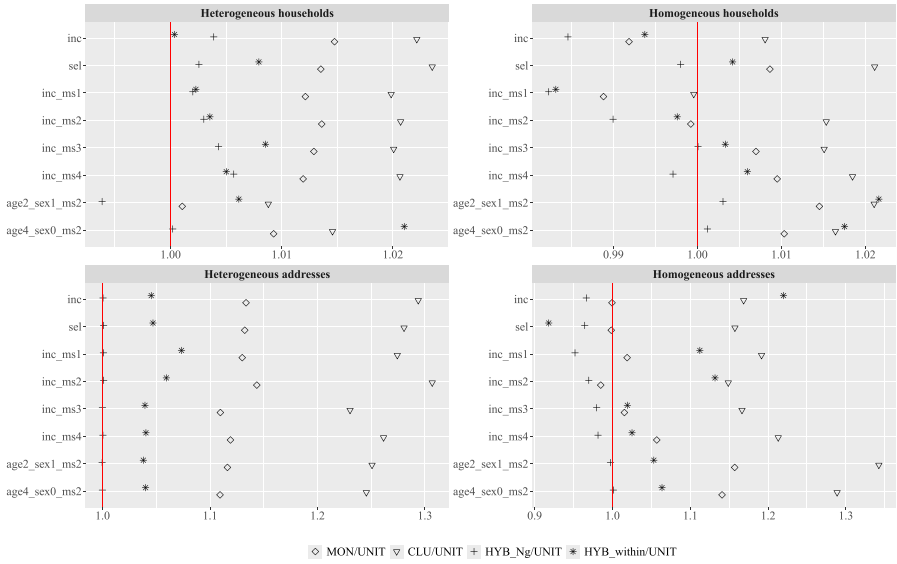

Fig. 3 Ratios of the MSE relative to the unit-level GREG estimator for small samples sizes ($m = 300$ clusters resulting $\bar{n} = 867$ units in the household scenarios and $m = 75$ clusters resulting in $\bar{n} = 817$ units in the address scenarios)

size. As quality criteria, we use the empirical mean squared error (MSE) defined by

$$\text{MSE}(\widehat{\tau}) = \frac{1}{10000} \sum_{r=1}^{10000} (\widehat{\tau}_r - \tau)^2$$

with $\widehat{\tau}_r$ as the total estimate for the r-th MC iteration.

Figures 3 and 4 depict the ratio of the MSE of MON, CLU and HYB relative to the MSE of UNIT for different sample sizes. First, in the household scenarios (with smaller cluster sizes), the differences between the estimators are very small. Second, when comparing the unit- and cluster-level estimators, it becomes evident that the unit-level estimators outperform the cluster-level ones in scenarios with smaller sample sizes (Figure 3). For large sample sizes and heterogeneous populations, the unit-level estimators continue to yield better results for most variables. This indicates that cluster heterogeneity plays an important role in determining whether modeling at the unit or cluster level is preferable. The advantage of the unit-level estimators is likely attributable to the information loss that occurs when unit-level variables are aggregated at the cluster level, as well as to the smaller sample size of clusters compared to units. However, for large sample sizes and homogeneous clusters, the cluster-level estimator, particularly MON, achieves the most efficient estimates (upper and lower right plots in Figure 4).

Third, MON outperforms the cluster-level estimator across all scenarios and sample sizes. The disadvantage of CLU is mainly attributable to the variance matrix $\boldsymbol{V}_c$. When $\boldsymbol{V}_c$ is excluded from CLU, both cluster-level estimators perform very similarly. This suggests that the additional information incorporated in MON from other cluster members, in terms of cross-cluster correlations, is not highly relevant for the MSE of MON. Moreover, the efficiency loss of MON compared to the unit-level estimators in heterogeneous populations is not solely driven by the design-balanced variables, which are likewise not included in

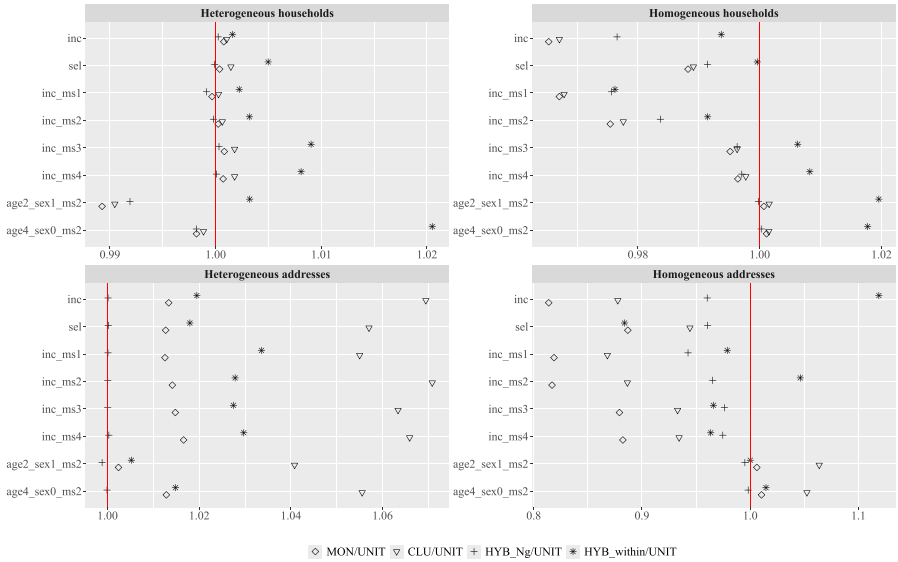

Fig. 4 Ratios of the MSE relative to the unit-level GREG estimator for large sample sizes ($m = 4.000$ clusters resulting in $\bar{n} = 11.560$ units in the household scenarios and $m = 1.000$ clusters resulting $\bar{n} = 10.882$ units in the address scenarios)

CLU. These results are remarkable because MON is the optimal estimator, minimizing the variance of the unit-level GREG estimator for large sample sizes. This suggests that, besides in addition to design-based variables for finite sample sizes, within-cluster heterogeneity also contribute to estimator instability, even for larger sample sizes. Fourth, HYB_Ng is more efficient than HYB_within in nearly all cases, indicating that the within-variance does not capture heterogeneity more effectively than the cluster size. Moreover, HYB_Ng has the additional advantage of producing multi-purpose weights.

In conclusion, for small cluster sizes, such as households, the differences between all estimators are minimal. However, with larger and variable cluster sizes, the differences become significant. In practice, clusters are typically larger and of variable size, as in regions, areas, schools, or other types of aggregated units, making the choice of estimator particularly important. Additionally, whether the population is homogeneous or heterogeneous depends on the study variables. Therefore, an estimator that performs well across the scenarios is advantageous. The hybrid estimator weighted by cluster sizes has the potential to do this. Independent from the sample size, in heterogeneous populations, it performs similarly to the unit-level estimator (or the differences are minimal as in the upper right plot in Figure 3). In homogeneous populations, the MSE of HYB_Ng is between the unit-level and cluster-level estimators.

The variance estimator of HYB is assessed with respect to the relative bias (RB) that measures the relative deviation of the estimated variance from the empirical variance of the point estimator. The RB of $\widehat{V}(\widehat{\tau})$ is defined by

$$\mathrm{RB}\left(\widehat{V}(\widehat{\tau})\right) = \frac{1}{10000} \sum_{r=1}^{10000} \frac{\widehat{V}(\widehat{\tau}_r) - V(\widehat{\tau})}{V(\widehat{\tau})}$$

with

$$V(\widehat{\tau}) = \frac{1}{10000} \sum_{r=1}^{10000} \left(\widehat{\tau}_r - \overline{\widehat{\tau}}\right)^2, \qquad \overline{\widehat{\tau}} = \frac{1}{10000} \sum_{r=1}^{10000} \widehat{\tau}_r$$

and $\widehat{V}(\widehat{\tau}_r)$ as the variance estimate for the r-th MC iteration. We refrain from tabulating the RB of the variance estimators because it was almost negligible in all scenarios except for the homogeneous addresses population and the small sample size, where the maximum absolute relative bias was below 0.02. These results confirm that (16) is the appropriate variance estimator for the hybrid GREG estimator.

6 Summary and Conclusion

In this paper, we proposed a hybrid GREG estimator for the estimation of unit-level characteristics. The name hybrid emphasizes that the proposed GREG estimator balances between unit- and cluster-level modeling. On the one hand, it is implemented at the unit level, which is the same level as the study variable and the assisting model. the hybrid estimator borrows information on the auxiliaries and study variable from the other cluster members in terms of intra-class correlations. By incorporating the information of the cluster members in addition to the unit-level information, the cluster sampling design is already accounted for in the point estimator. The receptive extent to which the information of the units and the cluster members is incorporated is determined by the weighting factors.

The simulation study demonstrates that the proposed hybrid GREG estimator serves as a compromise between the unit- and cluster-level modeling. Regardless of sample size and heterogeneity of the population, it performs at least as well as the unit- or cluster-level GREG estimators. In practice, one single estimator should be preferred to accommodate different heterogeneity levels within the clusters across different study variables. Future research should explore how the weighting factor can be optimized to improve the efficiency of the hybrid estimator, particularly in homogeneous populations with larger sample sizes.

The hybrid estimator can be easily extended to multi-stage cluster sampling designs. First, the inclusion probabilities of units and clusters differ. Second, further terms must be incorporated into the variance formula to capture the additional source of randomness introduced by each sampling stage.

[21] proposed a mixed model-assisted GREG estimator as a compromise between the Montanari estimator and the GREG estimator. This approach incorporates the sampling design in the random component of the model. Future research should compare the mixed model approach with the hybrid estimator in the context of cluster sampling.

A general drawback of the GREG estimator is that the weights can be negative. Reasons might be small sample sizes, a large number of auxiliary variables, or both. Most survey users strive to obtain positive weights. Fortunately, a variety of literature exists on how to restrict the range of the weights. First, the weights can be modified subsequently by using trimming methods [cf.1, 24]. Second, in the calibration estimator context bounds can be introduced to restrict the weights [cf.6, 11, 12, 14]. Thus, further research should address the question of how to embed the hybrid GREG estimator into the calibration estimator framework.

A relevant issue in cluster surveys is internal consistency, which means it is desirable to guarantee the same estimates for variables that are common to the unit- and cluster-level data set. Appropriate methods to guarantee internal consistency are integrated weighting [cf.16, 23] or the two-level GREG estimator [15]. Thus, for practical applications, combining the

hybrid GREG estimator with either integrated weighting or the two-level GREG estimator is an interesting future research field.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Beaumont, J.- F.: A new approach to weighting and inference in sample surveys. Biometrika **95**(3), 539–553 (2008)
2. Berger, Y.G., Tirari, M.E., Tillé, Y.: Towards optimal regression estimation in sample surveys. Australian & New Zealand Journal of Statistics **45**(3), 319–329 (2003)
3. Burgard, J.P., Kolb, J.- P., Merkle, H., Münnich, R.: Synthetic data for open and reproducible methodological research in social sciences and official statistics. AStA Wirtschafts- und Sozialstatistisches Archiv **11**(3–4), 233–244 (2017)
4. Cassel, C.- M., Särndal, C., Wretman, J.: Foundations of inference in survey sampling. Wiley, New York (1977)
5. Cassel, C.M., Särndal, C.E., Wretman, J.H.: Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika **63**(3), 615–620 (1976)
6. Deville, J.- C., Särndal, C.- E.: Calibration estimators in survey sampling. J. Am. Stat. Assoc. **87**(418), 376–382 (1992)
7. Guandalini, A., Tillé, Y.: Design-based estimators calibrated on estimated totals from multiple surveys. Int. Stat. Rev. **85**(2), 250–269 (2017)
8. Hansen, M., Hurwitz, W., Madow, W.:Sample survey methods and theory (Vol. I and II). New York: Wiley (1953)
9. Hidiroglou, M.: Double sampling. Surv. Methodol. **27**(2), 143–154 (2001)
10. Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**(260), 663–685 (1952)
11. Huang, E., Fuller, W.: Non-negative regression estimation in sample survey data, pp. 300–305. Proccedings Social Statitics Section, American Statistical Association (1978)
12. Husain, M.: Construction of regression weights for estimation in sample surveys unpublished M.S. Thesis. Ames, IowaIowa State University (1969)
13. Isaki, C.T., Fuller, W.A.: Survey design under the regression superpopulation model. J. Am. Stat. Assoc. **77**(377), 89–96 (1982)
14. Isaki, C.T., Tsay, J.H., Fuller, W.A.: Weighting sample data subject to independent controls. Surv. Methodol. **30**(1), 35–44 (2004)
15. Konrad, A., Burgard, J.P., Münnich, R.: A two-level GREG estimator for consistent estimation in household surveys. Int. Stat. Rev. **89**(3), 635–656 (2021)
16. Lemaître, G., Dufour, J.: An integrated method for weighting persons and families. Survey Methodology **13**, 199–207 (1987). (https://www150.statcan.gc.ca/n1/pub/12-001-x/1987002/article/14607-eng.pdf)
17. Lohr, S.L.: Sampling: Design and analysis (2.ed.ed.). Boston, MA: Cengage Learning (2010)
18. Montanari, G.: Post-sampling efficient QR-prediction in large-sample surveys. Int. Stat. Rev. **55**(2), 191–202 (1987)
19. Montanari, G.: On the regression estimation of finite population means. Surv. Methodol. **24**(1), 69–77 (1998)
20. Montanari, G., Ranalli, G.: Asymptotically efficient generalised regression estimators. Journal of Official Statistics **18**(4), 577 (2002)
21. Montanari, G., Ranalli, G.: A mixed model-assisted regression estimator that uses variables employed at the design stage. Stat. Methods Appl. **15**, 139–149 (2006)

22. Narain, R.: On sampling without replacement with varying probabilities. Journal of Indian Society of Agricultutral Statistics **3**, 169–175 (1951)
23. Nieuwenbroek, N.: An integrated method for weighting characteristics of persons and households using the linear regression estimator. Netherlands Central Bureau of Statistics, Department of Statistical Methods (1993)
24. Potter, F.J.: A study of procedures to identify and trim extreme sampling weights. Proceedings of the american statistical association, section on survey research methods (Vol. 225230) (1990)
25. Robinson, W.: Ecological correlations and the behavior of individuals. Am. Sociol. Rev. **15**(3), 351–357 (1950)
26. Särndal, C.E.: On π-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika **67**(3), 639–650 (1980)
27. Särndal, C- E., Swensson, B., Wretman, J.: Model assisted survey sampling. Springer Science & Business Media (1992)
28. Valliant, R., Dever, J.A., Kreuter, F.: Practical tools for designing and weighting survey samples. Springer (2013)
29. Wright, R.L.: Finite population sampling with multivariate auxiliary information. J. Am. Stat. Assoc. **78**(384), 879–884 (1983)