Statistica & Applicazioni - Vol. XVII, n. 2, 2019, pp. 201-235

HYBRID CALIBRATION METHODS FOR SMALL DOMAIN ESTIMATION

Risto Lehtonen*
Ari Veijanen*

SUMMARY

Hybrid calibration refers to an approach where techniques of classical calibration and more recent model-assisted calibration are combined for a joint calibration methodology. The classical calibration does not assume a model but uses the original auxiliary data as aggregates, whereas in model calibration, unit-level predictions from a model are used as pseudo auxiliary information. By combining these approaches we introduce hybrid methods, where aggregate data from different levels of the population are supplied to the model-free component and unit-level data are incorporated into the model-assisted component. The choice of the model depends on the type of the target variable. We use here linear and logistic mixed models. In the estimation for population subgroups or domains, the classical calibration fails when domain sample sizes become small. Our hybrid calibration methods were more accurate in small domains. In our studies, the basic model-assisted calibration was usually the best in accuracy, but the method requires population-level information on auxiliary variables in the model. The basic hybrid calibration method overcomes this restriction by including a model-free calibration component in the model-assisted calibration procedure. A new two-level hybrid calibration technique provides a further extension applicable for hierarchically structured populations. In this method, calibration in the model-free part is performed at a higher regional level, instead of the domain level. In our simulation experiments, the two-level hybrid calibration performed well: its accuracy and design bias were comparable to model calibration. The most stable weight distributions were obtained by the two-level method and Hájek type estimators developed in the paper.

Keywords: Auxiliary Information, Model-assisted Calibration, Mixed Models, Survey Weights, Design-based Simulation Experiments.
DOI: 10.26350/999999_000026
ISSN: 18246672 (print)

1. INTRODUCTION

Calibration techniques using auxiliary data provide powerful tools for estimating finite population parameters (Huang and Fuller, 1978; Deville and Särndal, 1992; Wu and Sitter, 2001; Montanari and Ranalli, 2005; Kott, 2009; Chen, Valliant, Elliott, 2018). The calibration approach formalized in Deville and Särndal (1992) reproduces the published official statistics of auxiliary variables from sample data (coherence property) using only aggregate-level auxiliary data. Classical calibration is

* Department of Statistics - University of Helsinki - Unioninkatu, 35, P.O. Box 18 - 00014 University of Helsinki, HELSINKI (e-mail: ✉ risto.lehtonen@helsinki.fi; ari.veijanen@gmail.com).

called *model-free* because it avoids explicit model statement, so the same calibration weights can be applied to the set of variables of interest (multi-purposiveness). Moreover, the estimates are nearly design unbiased (Särndal, 2007). These features have been appreciated for decades in official statistics.

Reliable statistics are also required for various population subgroups or *domains*. A typical example of a domain is a NUTS region in the European Union. The important sub-divisions of the population are often pre-specified in the sampling design by stratification. Sample sizes in the strata or *planned* domains can be determined large enough for reliable results with classical model-free methods. Estimation is then carried out independently in each stratum by using direct estimators that use observations of the target variable only from the given domain. In small domain estimation, the domain structures are not necessarily known in advance but emerge afterwards. Examples of ad hoc domains are various socio-economic breakdowns of the population. Sample sizes in such *unplanned* domains are random variates and can be small. For small domains, the classical calibration with direct estimators may fail to yield coefficients of variation small enough to be published. More advanced estimators are needed. For example, Estevao and Särndal (1999, 2004), Lehtonen and Veijanen (2009) and Hidiroglou and Estevao (2016) discuss model-free calibration for domain estimation.

Model-free calibration relies on an implicit linear functional relationship between the target variable and covariates. Wu and Sitter (2001) introduced *model calibration* for nonlinear relationships. Wu (2003) showed that the method has certain optimal properties. Model calibration allows flexible modelling, including generalized linear models. The method requires, however, unit-level auxiliary variable values for the population. By using the model, predicted values are computed for population elements, and the weighted sum of the predictions over the sample is calibrated to the sum of predictions in the population.

Recent developments in model calibration using for example nonparametric and semiparametric methods and LASSO techniques are presented for example in Breidt and Opsomer (2009), Rueda, Sànchez-Borrego, Arcos and Martìnez (2010), Wang and Wang (2011) and McConville, Breidt, Lee and Moisen (2017). Dagdoug, Goga and Haziza (2020) discuss calibration through random forests. Chandra and Chambers (2011) presented a model-based view to the method of Wu and Sitter. Model calibration for small domain estimation is discussed for example in Fabrizi, Salvati, Pratesi and Tzavidis (2014), Lehtonen and Veijanen (2012, 2016) and Morales, Rueda and Esteban (2018). Burgard, Münnich and Rupp (2019) presented a generalized calibration approach ensuring coherent estimates with small area constraints. Nearly design-unbiased methods that incorporate a model are often called *model-assisted*.

The built-in coherence property of model-free calibration does not hold for model-assisted calibration. We introduce *hybrid calibration* methods to overcome this restriction. In hybrid calibration, the coherence property is allowed for a set of auxiliary variables by a model-free calibration component in the calibration procedure, and the model-assisted calibration component provides efficiency gains. A multiple model calibration method of Montanari and Ranalli (2009) involves similar goals.

Basic hybrid calibration sometimes yields almost as unstable weights as classical

model-free calibration, because the model-free calibration part is defined at the domain level where small sample sizes can be encountered. Stability was improved, when we dropped the following commonly used benchmark condition: the sum of weights over the sample domain equals the known population size of the domain. We call the resulting estimators *Hájek type estimators*. In the new *two-level hybrid calibration* technique, model-assisted calibration operates at the original domain level, but model-free calibration is defined at a higher hierarchical level of the population, in effect providing larger domain sample sizes and stable estimates. As in Lehtonen and Veijanen (2012), observations outside the domain under study contribute to the domain estimate, either directly or via a model.

Hybrid calibration easily incorporates more auxiliary data than model calibration. For example, if unit-level values of powerful auxiliary variables are in the sample data set but are available only as domain-level aggregates in the population data, basic hybrid calibration can be used to utilize this information. The two-level variant can be used, if the aggregates are available at a higher level only. These properties extend the applicability of the hybrid calibration methods beyond classical calibration and model-assisted calibration.

Estimation of design variance for calibration has been addressed in the recent literature. Population level estimation is discussed in Särndal, Swensson and Wretman (1992), Deville and Särndal (1992), Deville (1999) and Kott (2009), and domain level estimation is discussed in Lehtonen, Särndal and Veijanen (2003), Hidiroglou and Patak (2004) and Lehtonen and Veijanen (2009). Torabi and Rao (2008) address MSE estimation of a GREG estimator assisted by a mixed model. Variance estimation in model calibration is discussed for example in Wu and Sitter (2001) and Kim and Park (2010), who developed a linearization variance estimator applicable for model calibration. Canty and Davison (1999) proposed linearized jackknife and bootstrap. Rueda, Arcos, Molina and Trujillo (2018) discussed linearization and jackknife variance estimators for model calibration.

We examined the design-based properties (bias and accuracy), variance estimation, and weight performance of the calibration methods introduced here by simulation experiments with a synthetic population and a real population obtained from statistical registers of Statistics Finland. Our main interest is in the domain totals and proportions in small domains. A *small domain* may be defined as a domain whose sample size is not large enough for acceptable precision with a direct estimator, such as the Horvitz-Thompson estimator. We modelled continuous variables by linear mixed models, and binary variables by logistic mixed models. Mixed effects were included in the models to account for the possible heterogeneity across the population domains. Because the domain structures were of unplanned type, realized sample sizes were sometimes very small. We report results for domain sizes as small as 12 elements in the synthetic population and less than 25 elements in the real population.

The paper is organized as follows. Notation and models are presented in Section 2. Calibration methods are introduced in Sections 3 and 4. Empirical results are presented in Section 5 for the synthetic population and in Section 6 for the real population. Conclusions are in Section 7.

2. NOTATION AND MODELS

Consider a unit-level finite population $U = \{1, 2, \ldots, k, \ldots, N\}$ of size N identifiable units, where k refers to the label of population element. Sub-populations of U or domains of interest are denoted U_d, $\cup_{d=1}^{D} U_d = U$, $d = 1, 2, \ldots, D$, and D is the number of domains. The size of domain U_d is N_d, $\sum_{d=1}^{D} N_d = N$. The domains are structured into domain-dependent higher-level sub-populations or regions (supersets) denoted $U_{r(d)}$ ($U_{r(d)} \supset U_d$).

A sample $s \subset U$ of n units is drawn from U with sampling design $p(\cdot)$ involving design weights denoted $a_k = 1/\pi_k$, where π_k is inclusion probability for element $k \in U$. The corresponding subsets of sample s are $s_d = U_d \cap s$ and $s_{r(d)} = U_{r(d)} \cap s$, $s_{r(d)} \supset s_d$. We assume that there are no empty sets s_d. Because the allocation of the sample s into domains U_d is not controlled by the sampling design, the realized domain sample sizes n_d are random variates. The domains thus are of unplanned type, and therefore, n_d can be small in some domains.

Auxiliary information of the population on variables related to the target variables of the survey plays a crucial role. We assume an access to unit-level auxiliary information; let $\mathbf{x}_k = (x_{1k}, \ldots, x_{Jk})'$ denote a vector value known for population element $k \in U$. In model-assisted methods, constant $x_{0k} = 1$ for all k is often inserted. Values y_k of the target variable are obtained for sample elements $k \in s$. Sample observations and auxiliary data are merged at the unit level by using unique identifiers from both data sources. This option gives flexibility and is available increasingly often in modern statistical infrastructures. We assume a complete data set without missingness.

For a continuous target variable, the parameters of interest are *domain totals*

$$t_d = \sum_{k \in U_d} y_k, \quad d = 1, \ldots, D, \tag{1}$$

and for a binary target variable they are *domain proportions*

$$p_d = \frac{t_d}{N_d} = \frac{\sum_{k \in U_d} y_k}{N_d}, \quad d = 1, \ldots, D, \tag{2}$$

where $y_k = 1$ refers to the occurrence of the event of interest and $y_k = 0$ otherwise.

Estimators for domain parameters (1) and (2) include various types of Horvitz-Thompson and Hájek estimators, which will be defined as direct or indirect estimators (Federal Committee on Statistical Methodology, 1993) and further as semi-direct or semi-indirect estimators (Lehtonen and Veijanen, 2012). Hidiroglou and Patak (2004) discuss Horvitz-Thompson and Hájek estimators for small domain estimation.

The possible heterogeneity across domains should be taken into account in small domain estimation. Fixed-effects models involving domain-specific or region-specific effects can become infeasible if the number of domains is large. We use members of the family of generalized linear mixed models (GLMM) with random effects defined at the domain level. Linear and logistic model formulations are natural choices when working with continuous and binary target variables. A *linear mixed model* with domain-specific random intercepts u_{0d} is given by

$$Y_k = \mathbf{x}_k'\beta + u_{0d} + \varepsilon_k,\ k \in U_d, u_{0d} \sim N(0, \sigma_u^2), \varepsilon_k \sim N(0, \sigma^2), \tag{3}$$

and a *logistic mixed model* is:

$$E_m(y_k|u_{0d};\beta) = P\{y_k = 1|u_{0d};\beta\} = \frac{\exp(\mathbf{x}_k'\beta + u_{0d})}{1 + \exp(\mathbf{x}_k'\beta + u_{0d})}, u_{0d} \sim N(0, \sigma_u^2), \tag{4}$$

where $k \in U_d$, $\mathbf{x}_k = (x_{0k}, x_{1k}, \ldots, x_{Jk})'$ is the vector of auxiliary variable values for element k, $\beta = (\beta_0, \beta_1, \ldots, \beta_J)'$ is a vector of fixed effects common for all domains and m refers to the expectation under the model. The parameters β, σ^2 and σ_u^2 are first estimated by maximum likelihood methods (e.g. R package *lme4* or SAS procedures MIXED and GLIMMIX), and the values $\hat{u}_{0d}$ of the random effects are calculated. Predictions $\hat{y}_k = P\{y_k = 1|\hat{u}_{0d};\hat{\beta}\}$ for the target variable values in the population are then computed for $k \in U_d$, $d = 1, \ldots, D$. The predicted y-values for the population constitute the key building blocks for the model-assisted methods. Lehtonen, Särndal and Veijanen (2003, 2005) give several special cases of models (3) and (4). Survey weights can be incorporated in the estimation equations to account for the possible informative sampling, for example. Morales *et al.* (2018) and Burgard and Dörr (2018) discuss some recent developments.

3. STANDARD ESTIMATORS

The ordinary Horvitz-Thompson estimator (Horvitz and Thompson, 1952) is often used for population totals and means. As a contribution to a discussion on the classical elephant example of Basu (1971), Jaroslav Hájek proposed an estimator for the population mean under PPS sampling as an alternative to the Horvitz-Thompson estimator, which was known to behave unexpectedly in certain situations. For example, Särndal *et al.* (1992, p. 182) have presented situations for preferring the Hájek estimator, such as if a) the differences between observations and population mean is small, or b) sample size is not fixed, or c) inclusion probabilities correlate weakly with the observations or the correlation is negative. We also adapt a Hájek estimator for our purposes. The variant developed here tends to provide more stable performance for small domains in particular.

A *Horvitz-Thompson* (HT) estimator for domain totals of a continuous target variable is given as

$$\hat{t}_d^{HT} = \sum_{k \in s_d} a_k y_k, \quad d = 1, \ldots, D, \tag{5}$$

where $a_k = 1/\pi_k$ are design weights. The HT estimator is of direct type as it only involves observations from the given domain. Estimator (5) does not involve auxiliary information. If domain sizes N_d are known, a direct *Hájek* (HA) estimator (Hájek, 1971) is of the form

$$\hat{t}_d^{HA} = N_d \left(\frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k} \right), \quad d = 1, \ldots, D, \tag{6}$$

and is often preferred (e.g. Rao, 1966).

Assuming known N_d, HT and Hájek type estimators of domain proportions of a binary variable are:

$$\hat{p}_d^{HT} = \frac{\hat{t}_d^{HT}}{N_d} = \frac{\sum_{k \in s_d} a_k y_k}{N_d}, \quad d = 1, \ldots, D,$$

and

$$\hat{p}_d^{HA} = \frac{\hat{t}_d^{HA}}{N_d} = \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k}, \quad d = 1, \ldots, D. \tag{7}$$

4. CALIBRATION ESTIMATORS

4.1 *Calibration in domain estimation*

Deville and Särndal (1992) presented the basic design-based model-free calibration technique. We discuss here a calibration weighting system for domain estimation by deriving calibration equations for a given calibration vector and solving the equations under a chi-square type distance function. In domain estimation, *calibration equations* are given by

$$\sum_{i \in s_d} w_{di} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i, \quad d = 1, \ldots, D, \tag{8}$$

where w_{di} is calibration weight for element i in domain d and $\mathbf{z}_i$ denotes a generic calibration vector whose structure depends on the chosen calibration method, defined for model-free calibration, model-assisted calibration and hybrid calibration in the next three sections. The new two-level hybrid calibration method involves level-specific calibration vectors and is discussed in Section 4.5.

We employ the distance measure approach with a chi-square distance (Deville and Särndal, 1992). Using Lagrange multipliers λ_d we minimize:

$$\sum_{k \in s_d} \frac{(w_{dk} - a_k)^2}{a_k} - \lambda_d' \left(\sum_{i \in s_d} w_{di} \mathbf{z}_i - \sum_{i \in U_d} \mathbf{z}_i \right), \quad d = 1, \ldots, D, \tag{9}$$

subject to the calibration conditions (8). The equation is minimized by weights

$$w_{dk} = a_k \left(1 + \lambda_d' \mathbf{z}_k\right), \tag{10}$$

where

$$\lambda_d' = \left(\sum_{i \in U_d} \mathbf{z}_i - \sum_{i \in s_d} a_i \mathbf{z}_i \right)' \left(\sum_{i \in s_d} a_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1}, \quad d = 1, \ldots, D. \tag{11}$$

We assume $\sum_{i\in s_d} a_i \mathbf{z}_i \mathbf{z}_i'$ is invertible. In domain estimation, the weights (10) are applied over a domain or region.

We introduce the term *Hájek type* for calibration with a calibration vector $\mathbf{z}_i$ that does not contain the constant 1 as the first element, as is the case for Horvitz-Thompson type estimators. Our Hájek estimators are simple generalizations of the ordinary Hájek estimators (6) and (7). In fact, a model calibration estimator incorporating an uninformative model with constant predictions is identical with the Hájek estimator.

The required calibration vectors for the specific calibration methods discussed in the next three sections are presented in Table 1.

TABLE 1. - *Calibration vectors for Horvitz-Thompson (HT) and Hájek (HA) type model-free (MFC), model-assisted (MC) and hybrid (HC) calibration estimators*

MFC	HT type: Calibration vector $\mathbf{z}_i = \left(1, \mathbf{x}_{Ci}'\right)'$, $i \in U_d$ HA type: Calibration vector $\mathbf{z}_i = \mathbf{x}_{Ci}$, $i \in U_d$ $\mathbf{x}_{Ci} = (x_{1i}, \ldots, x_{Ji})'$ calibration x-vector
MC	HT type: Calibration vector $\mathbf{z}_i = (1, \hat{y}_i)'$, $i \in U_d$ HA type: Calibration vector $\mathbf{z}_i = \hat{y}_i$, $i \in U_d$ $\hat{y}_i = f\left(\mathbf{x}_{Mi}'(\hat{\beta} + \hat{\mathbf{u}}_d)\right), i \in U_d$, predictions from the mixed model $\mathbf{x}_{Mi} = (1, x_{1i}, \ldots, x_{Ji})'$ model x-vector
HC	HT type: Calibration vector $\mathbf{z}_i = \left(1, \hat{y}_i, \mathbf{x}_{Ci}'\right)'$, $i \in U_d$ HA type: Calibration vector $\mathbf{z}_i = \left(\hat{y}_i, \mathbf{x}_{Ci}'\right)$, $i \in U_d$ $\hat{y}_i = f\left(\mathbf{x}_{Mi}'(\hat{\beta} + \hat{\mathbf{u}}_d)\right), i \in U_d$, predictions from the mixed model $\mathbf{x}_{Ci}$ calibration x-vector, $\mathbf{x}_{Mi}$ model x-vector $\mathbf{x}_{Ci}$ and $\mathbf{x}_{Mi}$ are separate or overlapping sub-vectors of $\mathbf{x}_i$

4.2 *Model-free calibration*

In classical model-free calibration (Deville and Särndal, 1992), a calibration equation is imposed: the weighted sample sums of auxiliary x-variable values reproduce the known population sums (coherence or benchmarking property). The MFC calibration vector $\mathbf{z}_i$ for (10) contains the original auxiliary x-variables and is given for the HT and HA type estimators in Table 1.

Calibration equations (8) for HT type estimators is given by

$$\sum_{i\in s_d} w_{di}^{HT} \mathbf{z}_i = \sum_{i\in U_d} \mathbf{z}_i = \left(N_d, \sum_{i\in U_d} x_{1i}, \ldots, \sum_{i\in U_d} x_{Ji}\right)', \quad d = 1, \ldots, D. \tag{12}$$

For HT type MFC estimator of domain totals (1) we minimize (9) subject to (12) and obtain

$$\hat{t}_{dMFC}^{HT} = \sum_{k\in s_d} w_{dk}^{HT} y_k, \quad d = 1, \ldots, D, \tag{13}$$

where MFC weights w_{dk}^{HT} are computed by (10) and (11) with z-vector $\mathbf{z}_i = (1, x_{1i}, \ldots, x_{Ji})'$.

Calibration of (8) for Hájek type estimators is:

$$\sum_{i \in s_d} w_{di}^{HA} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i = \left(\sum_{i \in U_d} x_{1i}, \ldots, \sum_{i \in U_d} x_{Ji} \right)', \quad d = 1, \ldots, D. \tag{14}$$

For Hájek type MFC estimator of domain totals (1) we minimize (9) subject to (14) and obtain:

$$\hat{t}_{dMFC}^{HA} = N_d \frac{\sum_{k \in s_d} w_{dk}^{HA} y_k}{\sum_{k \in s_d} w_{dk}^{HA}}, \quad d = 1, \ldots, D, \tag{15}$$

where MFC weights w_{dk}^{HA} are computed by (10) and (11) with z-vector $\mathbf{z}_i = (x_{1i}, \ldots, x_{Ji})'$.

Using again the MFC calibration weights, HT and Hájek type estimators of domain proportions (2) of a binary variable are derived as

$$\hat{p}_{dMFC}^{HT} = \frac{\hat{t}_{dMFC}^{HT}}{N_d} = \frac{\sum_{k \in s_d} w_{dk}^{HT} y_k}{N_d}, \quad d = 1, \ldots, D, \tag{16}$$

and

$$\hat{p}_{dMFC}^{HA} = \frac{\hat{t}_{dMFC}^{HA}}{N_d} = \frac{\sum_{k \in s_d} w_{dk}^{HA} y_k}{\sum_{k \in s_d} w_{dk}^{HA}}, \quad d = 1, \ldots, D. \tag{17}$$

Estimators (13) and (15) - (17) are of direct type because they use y_k values from the given domain only. Under the setting described in Section 2, it may happen that a small number of sample elements only fall in some domains, possibly causing instability problems for a MFC estimator in such domains.

4.3 *Model-assisted calibration*

Model-assisted calibration (MC) for domain estimation extends classical calibration techniques by introducing modelling in the calibration procedure. In addition to an expected improvement in accuracy over MFC, the method aims to overcome the possible instability problems of MFC in small domains. In the *modelling step* of a MC procedure, the chosen model is fitted to the entire sample data set, and predictions $\hat{y}_k$ are computed for every $k \in U_d$, $d = 1, \ldots, D$, by using unit-level auxiliary x-data from the population. In the *calibration step*, the predictions are incorporated in the calibration z-vector and calibration weights are determined by using the machinery presented in Section 4.1.

Model-assisted calibration can be defined at various hierarchical levels of the population, including population level and domain level as well as intermediate levels,

for example a neighbourhood that contains the domain of interest (Lehtonen and Veijanen, 2012). In a *semi-direct* approach, the calibration step only involves y-values and predictions for the domain of interest. However, the MC method discussed in this section and the version of hybrid calibration of Section 4.4 are not strictly direct methods because predictions are computed by a model fitted to the entire sample. Lehtonen and Veijanen (2012) presented an extension of MC that was defined as *semi-indirect*. The two-level hybrid calibration method developed in Section 4.5 is *indirect*, as observed y-values and predictions outside the domain of interest also contribute.

Wu and Sitter (2001) introduced model calibration for population level calibration. The calibration weights must satisfy calibration equation

$$\sum_{i\in s} w_i \mathbf{z}_i = \sum_{i\in U} \mathbf{z}_i = \left(N, \sum_{i\in U} \hat{y}_i\right)',$$

where $\mathbf{z}_i = (1, \hat{y}_i)'$ and $\hat{y}_i$ are predictions from the model. The selection of $\mathbf{z}_i$ is not unique; Wu and Sitter (2001) consider different options. We discuss here semi-direct calibration for domain estimation. The HT and Hájek type MC calibration vectors $\mathbf{z}_i$ for (9) were given in Table 1. Both z-vectors involve fitted y-values; they are computed as

$$\hat{y}_k = f(\mathbf{x}'_{Mk}(\hat{\beta} + \hat{\mathbf{u}}_d)),\ k \in U_d, \quad d = 1, \ldots, D,$$

with the respective model x-vectors $\mathbf{x}_{Mk}$, $k \in U_d$, where f refers to the chosen link function, $\hat{\beta}$ is the vector of estimated fixed effects common for all domains and $\hat{\mathbf{u}}_d$ are vectors of predicted random effects (intercepts and slopes). For a continuous variable, we use a linear link function and obtain predictions from a linear mixed model (3) with domain-specific random intercepts $\hat{u}_{0d}$ as given by

$$\hat{y}_k = \mathbf{x}'_{Mk}\hat{\beta} + \hat{u}_{0d},\ k \in U_d, \quad d = 1, \ldots, D. \tag{18}$$

For a binary variable we compute the predictions from a logistic mixed model (4):

$$\hat{y}_k = \frac{\exp(\mathbf{x}'_{Mk}\hat{\beta} + \hat{u}_{0d})}{1 + \exp(\mathbf{x}'_{Mk}\hat{\beta} + \hat{u}_{0d})},\ k \in U_d, \quad d = 1, \ldots, D. \tag{19}$$

HT type MC calibration equations are given by

$$\sum_{i\in s_d} w_{di}^{HT} \mathbf{z}_i = \sum_{i\in U_d} \mathbf{z}_i = \left(N_d, \sum_{i\in U_d} \hat{y}_i\right)', \quad d = 1, \ldots, D, \tag{20}$$

where predictions for continuous y-variable are computed by (18) and for binary variable by (19).

For model-assisted HT type MC calibration estimator of totals (1) we minimize (9) subject to (20) using predictions from (18) and obtain

$$\hat{t}_{dMC}^{HT} = \sum_{k \in s_d} w_{dk}^{HT} y_k, \quad d = 1, \ldots, D, \tag{21}$$

where weights w_{dk}^{HT} are computed by (10) and (11) with z-vector $\mathbf{z}_i = (1, \hat{y}_i)'$.

Hájek type MC calibration equations are given by

$$\sum_{i \in s_d} w_{di}^{HA} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i = \sum_{i \in U_d} \hat{y}_i, \quad d = 1, \ldots, D, \tag{22}$$

where predictions for continuous y-variable are computed by (18) and for binary variable by (19).

For model-assisted Hájek type MC calibration estimator of totals (1) we minimize (9) subject to (22) using predictions from (18) and obtain

$$\hat{t}_{dMC}^{HA} = N_d \frac{\sum_{k \in s_d} w_{dk}^{HA} y_k}{\sum_{k \in s_d} w_{dk}^{HA}}, \quad d = 1, \ldots, D, \tag{23}$$

where weights w_{dk}^{HA} are computed by (10) and (11) with z-vector $\mathbf{z}_i = \hat{y}_i$.

For a model-assisted HT type MC calibration estimator of proportions (2) we first estimate the domain total of the binary response variable using predictions (19). We then divide the estimated domain total $\hat{t}_{dMC}^{HT}$ (21) by the known domain size N_d:

$$\hat{p}_{dMC}^{HT} = \frac{\hat{t}_{dMC}^{HT}}{N_d}, \quad d = 1, \ldots, D. \tag{24}$$

The Hájek type counterpart is derived from $\hat{t}_{dMC}^{HA}$ (23):

$$\hat{p}_{dMC}^{HA} = \frac{\hat{t}_{dMC}^{HA}}{N_d}, \quad d = 1, \ldots, D. \tag{25}$$

4.4 *Hybrid calibration*

The coherence properties (12) and (14) of model-free calibration for the x-variable totals is lost in model-assisted calibration. In hybrid calibration, we impose the coherence property for a chosen subset of x-variables of the z-vector (the MFC part) and retain the MC calibration properties (20) and (22) for predictions computed for another subset (the MC part). Hybrid calibration is discussed here for a semi-direct HC, where calibration is applied at the domain level but the weights are computed for a model fitted to the entire sample.

For deriving HC calibration vectors, we split the original J element auxiliary x-vector into two nonoverlapping subsets: MFC part with calibration x-vectors $\mathbf{x}_{Ci} = \left(x_{1i}, \ldots, x_{ji}\right)'$ and MC part with model x-vectors $\mathbf{x}_{Mi} = \left(1, x_{j+1,i}, \ldots, x_{Ji}\right)'$. It also is possible to apply overlapping decomposition of the x-vectors, as in Section 5 example; the former way is applied in Section 6.

Calibration vectors for HT type hybrid calibration (HC) are:

$$\mathbf{z}_i = \left(1, \hat{y}_i, \mathbf{x}_{Ci}'\right)', \quad i \in U_d, d = 1, \ldots, D, \tag{26}$$

where $\mathbf{x}_{Ci} = \left(x_{1i}, \ldots, x_{ji}\right)'$. Calibration equations (8) are:

$$\sum_{i \in s_d} w_{di}^{HT} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i = \left(N_d, \sum_{i \in U_d} \hat{y}_i, \sum_{i \in U_d} x_{1i}, \ldots, \sum_{i \in U_d} x_{ji} \right)', \tag{27}$$

where predictions for continuous y-variable are computed by (18) and for binary variable by (19) by using model x-vectors $\mathbf{x}_{Mk} = \left(1, x_{j+1,k}, \ldots, x_{Jk}\right)'$.

We minimize (9) subject to (27) and obtain HT type HC estimator of totals (1):

$$\hat{t}_{dHC}^{HT} = \sum_{k \in s_d} w_{dk}^{HT} y_k, \quad d = 1, \ldots, D, \tag{28}$$

where w_{dk}^{HT} are computed by (10) and (11) with z-vector (26). Fitted values for the MC part are $\hat{y}_k = f(\mathbf{x}'_{Mk}(\hat{\beta} + \hat{\mathbf{u}}_d))$ with model x-vector $\mathbf{x}_{Mk} = (1, x_{j+1,k}, \ldots, x_{Jk})'$, $k \in U_d$, where the assisting model is (3).

Calibration vectors for Hájek type hybrid calibration are:

$$\mathbf{z}_i = \left(\hat{y}_i, \mathbf{x}'_{Ci}\right)', i \in U_d, \quad d = 1, \ldots, D, \tag{29}$$

where again predictions for continuous y-variable are computed by (18) and for binary variable by (19) with calibration x-vector $\mathbf{x}_{Ci}$ and model x-vector $\mathbf{x}_{Mk}$ as in (26) and (28). Calibration of (8) is:

$$\sum_{i \in s_d} w_{di}^{HA} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i = \left(\sum_{i \in U_d} \hat{y}_i, \sum_{i \in U_d} x_{1i}, \ldots, \sum_{i \in U_d} x_{ji} \right)'. \tag{30}$$

We minimize (9) subject to (30) and obtain HA type hybrid calibration estimator of totals (1):

$$\hat{t}_{dHC}^{HA} = N_d \frac{\sum_{k \in s_d} w_{dk}^{HA} y_k}{\sum_{k \in s_d} w_{dk}^{HA}}, \quad d = 1, \ldots, D, \tag{31}$$

where w_{dk}^{HA} are computed by (10) and (11) with z-vector (29).

An HC estimator of a domain proportion (2) is derived from the corresponding estimator of a domain total (1) that incorporates predictions (19). From the HT type HC estimator (28) of the domain total we obtain

$$\hat{p}_{dHC}^{HT} = \frac{\hat{t}_{dHC}^{HT}}{N_d}, \quad d = 1, \ldots, D. \tag{32}$$

The Hájek type HC estimator of a domain proportion incorporates (31):

$$\hat{p}_{dHC}^{HA} = \frac{\hat{t}_{dHC}^{HA}}{N_d}, \quad d = 1, \ldots, D. \tag{33}$$

4.5 *Two-level hybrid calibration*

The MFC part of hybrid calibration can involve instability for domains whose sample sizes are small. A new two-level hybrid calibration estimator is intended to reduce the effects of the possible instability but at the same time to retain the options for the coherence property (now, at the region level) and the efficiency improvement property of the MC method for the domain estimates. Two-level hybrid calibration (HC2) provides an indirect domain estimation method, because *y*-values from outside the domain are included; HC2 thus *borrows strength* from higher-level regions.

In two-level hybrid calibration, the model-assisted calibration part of the HC2 estimator is assigned to the original domain level (e.g. LAU-1), and the model-free calibration part is defined at a higher hierarchical level (e.g. NUTS-3), where instability problems might not be encountered. We present the HC2 method both for HT and Hájek type estimators for domain totals and proportions. Calibration vectors for the HC2 variants are presented in Table 2.

TABLE 2. - *Calibration vectors for Horvitz-Thompson (HT) and Hájek (HA) type two-level hybrid (HC2) calibration estimators*

Level 1 (domains) calibration vectors for MC part,
HT type: $\mathbf{z}_i^{(1)} = \left(x_{0i}^{(1)}, \hat{y}_i^{(1)}\right)', i \in U_{r(d)}$
HA type: $\mathbf{z}_i^{(1)} = \hat{y}_i^{(1)}, i \in U_{r(d)}$
$x_{0i}^{(1)} = 1, \hat{y}_i^{(1)} = \hat{y}_i, i \in U_d$
$x_{0i}^{(1)} = 0, \hat{y}_i^{(1)} = 0, i \in U_{r(d)} \setminus U_d$
$\hat{y}_i = f\left(\mathbf{x}'_{Mi}(\hat{\beta} + \hat{\mathbf{u}}_d)\right), i \in U_d, \mathbf{x}_{Mi}$ model *x*-vector
Level 2 (regions) calibration vectors for MFC part
$\mathbf{z}_i^{(2)} = \mathbf{x}_{Ci}, i \in U_{r(d)}, \mathbf{x}_{Ci}$ calibration *x*-vector
$\mathbf{x}_{Ci}$ and $\mathbf{x}_{Mi}$ are separate or overlapping sub-vectors of $\mathbf{x}_i$

In HT and Hájek type two-level hybrid calibration, we need to define two sets of calibration equations to be satisfied simultaneously. Calibration equations for Horvitz-Thompson type estimators in the separate calibration and model *x*-vectors case are:

$$\sum_{i \in s_{r(d)}} w_{r(d),i}^{HT} \mathbf{z}_i^{(1)} = \sum_{i \in U_d} \mathbf{z}_i^{(1)} = \left(\sum_{i \in U_d} x_{0i}^{(1)}, \sum_{i \in U_d} \hat{y}_i \right)' \text{ (MC part)} \tag{34}$$

and

$$\sum_{i \in s_{r(d)}} w_{r(d),i}^{HT} \mathbf{z}_i^{(2)} = \sum_{i \in U_{r(d)}} \mathbf{z}_i^{(2)} = \left(\sum_{i \in U_{r(d)}} x_{1i}, \ldots, \sum_{i \in U_{r(d)}} x_{ji} \right)' \text{(MFC part)}, \tag{35}$$

where

$s_{r(d)} = s \cap U_{r(d)}, U_{r(d)} \supset U_d, d = 1, \ldots, D,$

$\mathbf{z}_i^{(1)} = (x_{0i}^{(1)}, \hat{y}_i^{(1)}), i \in U_{r(d)}, d = 1, \ldots, D$ (level 1 calibration vector, MC part),

$x_{0i}^{(1)} = 1, i \in U_d, 0$ otherwise, (extended x_0-variable),

$\hat{y}_i^{(1)} = \hat{y}_i, i \in U_d, 0$ otherwise, (extended predictions),

$\hat{y}_i = f(\mathbf{x}'_{Mi}(\hat{\beta} + \hat{\mathbf{u}}_d)), i \in U_d$ (the chosen GLMM),

$\mathbf{x}_{Mi} = (1, x_{j+1,i}, \ldots, x_{Ji})', i \in U_d$ (model x-vector for MC part),

$\mathbf{z}_i^{(2)} = \mathbf{x}_{Ci} = (x_{1i}, \ldots, x_{ji})', i \in U_{r(d)}$ (level 2 calibration x-vector, MFC part).

Using Lagrange multipliers $\lambda'_{r(d)} = (\lambda'_1, \lambda'_2)$ we minimize:

$$\sum_{k \in s_{r(d)}} \frac{(w_{r(d),k}^{HT} - a_k)^2}{a_k} - \lambda'_{r(d)} \left(\sum_{i \in s_{r(d)}} w_{r(d),i}^{HT} \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix} - \begin{pmatrix} \sum_{i \in U_{r(d)}} \mathbf{z}_i^{(1)} \\ \sum_{i \in U_{r(d)}} \mathbf{z}_i^{(2)} \end{pmatrix} \right) \tag{36}$$

subject to calibration constraints (34) and (35). Writing $\mathbf{z} = \left(\mathbf{z}_k^{(1)'}, \mathbf{z}_k^{(2)'} \right)'$, Equation (36) is minimized by weights

$$w_{r(d),k}^{HT} = a_k \left(1 + \lambda'_{r(d)} \mathbf{z}_k \right),$$

where

$$\lambda'_{r(d)} = \left(\sum_{i \in U_{r(d)}} \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix} - \sum_{i \in s_{r(d)}} a_i \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix} \right)' \left(\sum_{i \in s_{r(d)}} a_i \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix} \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix}' \right)^{-1}.$$

The resulting HT type two-level HC estimator of domain total (1) is given by

$$\hat{t}_{dHC2}^{HT} = \sum_{k \in s_{r(d)}} w_{r(d),k}^{HT} y_k, \quad d = 1, \ldots, D.$$

The Hájek type counterpart is obtained by removing $\sum_{i \in U_d} x_{0i}^{(1)}$ from (34) and $x_{0i}^{(1)}$ from vector $\mathbf{z}_i^{(1)}$ (Table 2). A Hájek type HC2 estimator of domain total is then given by

$$\hat{t}_{dHC2}^{HA} = N_d \left(\frac{\sum_{k \in s_{r(d)}} w_{r(d),k}^{HA} y_k}{\sum_{k \in s_{r(d)}} w_{r(d),k}^{HA}} \right), \quad d = 1, \ldots, D.$$

The HT and Hájek type HC2 estimators of domain proportions are obtained by

$$\hat{p}_{dHC2}^{HT} = \frac{\hat{t}_{dHC2}^{HT}}{N_d}, \quad d = 1, \ldots, D, \tag{37}$$

and

$$\hat{p}_{dHC2}^{HA} = \frac{\hat{t}_{dHC2}^{HA}}{N_d}, \quad d = 1, \ldots, D, \tag{38}$$

respectively. It is expected that weights $w_{r(d),k}$ for $k \in s_r(d)$ outside s_d are small in absolute value.

In simulation studies presented in Sections 5 and 6 we verify that the indirect nature of two-level hybrid calibration estimators for domain totals does not introduce bias in the domain estimates.

4.6 *Variance estimation*

As the domain estimators are nearly design unbiased, an estimator of variance probably suffices for the estimation of MSE. Kott (2009, p. 71) presents a variance estimator that we modified for HT type domain total estimators as follows:

$$v(\hat{t}_d) = \frac{n_d}{n_d - 1}\left(\sum_{k \in s_d} w_{dk}^2 e_k^2 - \frac{1}{n_d}\left(\sum_{k \in s_d} w_{dk} e_k\right)^2\right), \qquad d = 1, \ldots, D, \tag{39}$$

where $e_k = y_k - \hat{y}_k$ are residuals from the model. For the two-level estimator, the sums involve units in the enclosing region:

$$v(\hat{t}_{dHC2}) = \frac{n_d}{n_d - 1}\left(\sum_{k \in s_{r(d)}} w_{r(d),k}^2 e_k^2 - \frac{1}{n_d}\left(\sum_{k \in s_{r(d)}} w_{r(d),k} e_k\right)^2\right). \tag{40}$$

These equations yielded more accurate results than the estimator in Kim and Park (2010; equation 20).

5. SIMULATION EXPERIMENTS WITH SYNTHETIC DATA

5.1 *Accuracy comparison*

We examine empirically the accuracy of the various calibration estimators discussed in Sections 3 and 4 by using design-based Monte Carlo experiments applied to an artificially generated population. The population consists of one million units distributed among $D = 40$ domains U_d of different sizes. The domains are grouped into four regions $U_r(d)$, each comprising ten domains. There are 15 small domains ($N_d = 6000$), 15 medium-sized domains ($N_d = 20000$) and 10 large domains ($N_d = 61000$). The first region incorporates ten small domains ($N_1(d) = 60000$). The second region contains five small domains and five medium-sized domains ($N_2(d) = 130000$). The third region has ten medium-sized domains ($N_3(d) = 200000$). All the large domains are in the fourth region ($N_4(d) = 610000$).

The expected value of the target variable depends on the region $U_r(d)$, the domain d and two x-variables as follows:

$$y_k = 5(r(d) - 1.5) + u_{0d} + (3 + u_{1d})x_{1k} + (1 + u_{2d})x_{2k} + \varepsilon_k,$$

where $k \in u_{0d} \subset U_r(d)$, $r(d) = 1, 2, 3, 4$, $d = 1, \ldots, 40$. Random intercept u_{0d} has variance $\sigma^2_{u0} = 2$ and both random slopes have variance $\sigma^2_{u1} = \sigma^2_{u2} = 0.125$. The random slopes are correlated with the random intercept: $corr(u_{0d}, u_{1d}) = corr(u_{0d}, u_{2d}) = -0.5$. The error term ε_k is distributed as $N(0, 25)$.

The x-variables are independently normally distributed with variance $\sigma^2_{x1} = \sigma^2_{x2} = \sigma^2_{x3} = 1$. Their expected values depend on domain as follows: $E(x_{1k}) = 10 + (d_k - 1)/10$, $E(x_{2k}) = 10 - (d_k - 1)/10$ and $E(x_{3k}) = 0$, where $d_k = d$, $k \in U_d$. For the full model the intra-class correlation (as computed with the R function icc) amounts to about 64% (adjusted icc) or 53% (conditional icc).

In the constructed population, the variables x_1 and x_2 are negatively correlated with $corr(x_1, x_2) = -0.47$. The study variable y has variance $\sigma^2_y = 92.3$. The correlations of y with x-variables are $corr(y, x_1) = 0.66$, $corr(y, x_2) = -0.37$ and $corr(y, x_3) = 0$. Domain means of the y-variable were 39.3 for small domains, 49.0 for medium domains, and 58.2 for large domains.

With this setup it was possible to derive assisting models and calibration and model x-vectors of varying explanatory power. The basic linear mixed model $y_k = \beta_0 + u_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \varepsilon_k$ was used for the MC, HC and HC2 methods. The model x-vector thus contained all three x-variables. Note that the model is wrongly specified (random slopes are missing). The MFC method does not involve an explicit assisting model; all three x-variables were incorporated in the calibration x-vector. Sub-vectors were inserted in the calibration x-vectors for HC and HC2. Note that the share of the x-variables between model and calibration x-vectors is of *overlapping* type.

In the experiments, $K = 10000$ independent simple random samples without replacement (SRSWOR) were drawn from the population with sample size $n = 2000$. Because the domains were of unplanned type, elements were randomly allocated into the domains. Thus, a domain with small realized sample size may contain fewer units than there are calibration equations, causing that the matrix in Eq. (11) has no inverse. In simulations, only 0.01% of the estimates were ignored for this reason, irrespective of the method. The problem appeared in the minor domains class only.

Design bias and accuracy of estimators for domain totals and proportions were measured by absolute relative bias (ARB) and relative root mean squared error (RRMSE):

$$ARB\left(\hat{\theta}_d\right) = \frac{1}{\theta_d}\left|\frac{1}{K}\sum_{j=1}^{K}\hat{\theta}_d(s_j) - \theta_d\right| \tag{41}$$

$$RRMSE\left(\hat{\theta}_d\right) = \frac{1}{\theta_d}\sqrt{\frac{1}{K}\sum_{j=1}^{K}\left(\hat{\theta}_d(s_j) - \theta_d\right)^2}, \quad d = 1, \ldots, D, \tag{42}$$

where $\hat{\theta}_d(s_j)$ is the estimate from sample s_j for domain d, θ_d is the known parameter

value in domain d, and K is the number of simulated samples. We computed median ARB and median RRMSE over three domain size classes defined by expected domain sample size. The analysis design for the simulation experiments is depicted in Table 3.

TABLE 3. - *Calibration vectors for Horvitz-Thompson (HT) and Hájek (HA) type estimators*

Model-free calibration MFC	HT type: $\mathbf{z}_i = (1, x_{1i}, x_{2i}, x_{3i})', i \in U_d$ HA type: $\mathbf{z}_i = (x_{1i}, x_{2i}, x_{3i})', i \in U_d$
Model-assisted calibration MC	HT type: $\mathbf{z}_i = (1, \hat{y}_i)', i \in U_d$ HA type: $\mathbf{z}_i = \hat{y}_i, i \in U_d$
Hybrid calibration HC	HT type: $\mathbf{z}_i = (1, \hat{y}_i, x_{3i})', i \in U_d$ HA type: $\mathbf{z}_i = (\hat{y}_i, x_{3i})', i \in U_d$
Two-level hybrid calibration HC2	HT Level 1: $\mathbf{z}_i^{(1)} = (x_{0i}^{(1)}, \hat{y}_i^{(1)})', i \in U_{r(d)}$ HA Level 1: $\mathbf{z}_i^{(1)} = \hat{y}_i^{(1)}, i \in U_{r(d)}$ $x_{0i}^{(1)} = 1, \hat{y}_i^{(1)} = \hat{y}_i, i \in U_d$ $x_{0i}^{(1)} = 0, \hat{y}_i^{(1)} = 0, i \in U_{r(d)} \setminus U_d$ HT and HA Level 2: $\mathbf{z}_i^{(2)} = x_{3i}, i \in U_{r(d)}$
Predictions for HT and HA type MC, HC, HC2	$\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{u}}_{0d} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}, i \in U_d$

Our interest is in the relative accuracy of the calibration methods, when essentially the same auxiliary information is incorporated in the estimation procedure by using different methods: the supply of raw auxiliary data (as in MFC), model predictions (as in MC), or a mixture of these (as in HC and HC2). The main concern is in the possible accuracy differences, when the domain sample sizes become small. It was also important for us to investigate whether the new two-level calibration method can address the potential instability of the MFC method in small areas. The HC method was also important in this regard, because in addition to a MC part it involves a MFC part at the domain level. Note that in Table 3, the model x-vector contains always all three x-variables, and the third x-variable serves as the sole auxiliary variable in HC and HC2, but at different levels. Results on accuracy are in Table 4.

In both Horvitz-Thompson and Hájek type methods, the model-assisted estimators show similar overall accuracy. Model-free calibration behaves worst, as expected, and the difference to model-assisted methods is significant. The accuracy of the methods appears quite similar in the major domains and medium-sized domains. Largest differences are found in the group of minor domains. All model-assisted methods clearly outperform the classical model-free calibration. Model-assisted calibration was expected to show good accuracy in small domains; this is indeed the case, in the HT and Hájek methods. In hybrid calibration, a MFC part was included to attain co-

herence of auxiliary variables with published statistics. HC is, however, less accurate than MC, and is worst among the model-assisted methods in both families. This is probably due to instability in the MFC part. The two-level hybrid calibration, where the MFC part is defined at a higher regional level, successfully stabilizes HC. HC2 even indicates slightly better accuracy than MC in the HT group, and is the best in accuracy of all methods considered. For medium-sized and major domains the overall accuracy of HT type methods is better than that of Hájek type methods, for all estimators. Differences in small domains are small, except for MFC that shows better accuracy for the Hájek type method. Model-assisted HT type methods outperform all model-assisted Hájek type methods in small domains.

TABLE 4. - *Median root mean squared error RRMSE (%) of Horvitz-Thompson and Hájek type calibration estimators for domain totals in three domain sample size classes (synthetic population)*

Method	**Horvitz-Thompson type**				**Hájek type**			
	Expected domain sample size (in parentheses)				Expected domain sample size (in parentheses)			
	Minor (12)	Medium (40)	Major (122)	All	Minor (12)	Medium (40)	Major (122)	All
Model-free estimator								
MFC	8.82	1.62	0.78	1.72	6.39	1.89	0.91	1.98
Model-assisted estimators								
MC	4.29	1.58	0.78	1.67	4.53	1.85	0.91	1.96
HC	5.49	1.60	0.78	1.69	4.90	1.88	0.92	1.99
HC2	4.25	1.58	0.78	1.67	4.55	1.86	0.91	1.96

We studied MSE estimation with (39) and (40) using 10000 simulated samples from the synthetic population. Estimates for a small domain with less than three units were omitted from all the calculations. In the absence of an analytically derived MSE, we approximated the MSE of an estimator by the empirical mean squared error. Bias of the MSE estimator ψ was measured by relative bias (RB) calculated using empirical MSE ψ:

$$RB(\hat{\psi}_d) = B(\hat{\psi}_d)/\psi_d = \frac{1}{K}\sum_{j=1}^{K}\Big(\hat{\psi}_d(s_j) - \psi_d\Big)/\psi_d. \tag{43}$$

The MSE estimator (39) performed well in the case of model calibration (MC) (Tables 5 and 6). For the basic hybrid calibration method (HC), MSE estimation was unreliable in the small domains. The MSE estimates obtained by (40) for the two-level estimator (HC2) were comparable in accuracy to the MSE estimates for MC. De-

sign bias of MSE estimator for MC was negligible. In HC2, negative bias was present especially in small domains, where a large proportion of the calibrated weights are assigned to units in the enclosing region.

TABLE 5. - *Mean relative bias (%) of the MSE estimator in* 10000 *samples from the synthetic population. Minimum size of the sample for a domain is* 3 *units*

Method	*Expected domain sample size*			
	Minor (12)	Medium (40)	Major (122)	All
MC	−1.28	0.28	−0.51	−0.50
HC	8.29	0.34	−0.42	3.13
HC2	−8.51	−2.00	−1.20	−4.24

TABLE 6. - *Mean RRMSE (%) of the MSE estimator in* 10000 *samples from the synthetic population. Minimum size of the sample for a domain is* 3 *units*

Method	*Expected domain sample size*			
	Minor (12)	Medium (40)	Major (122)	All
MC	106	30.9	16.1	55.4
HC	3896	32.8	16.3	1477
HC2	102	30.0	15.9	53.3

5.2 *Distribution of weights*

The linear calibration approach chosen here can involve large variation of weights and negative weights that are often considered unfeasible in practical applications. Large variation and negative weights are expected in small domains in particular. We were interested in finding out if there are differences in the weight behaviour between the methods, and to what extent (if any) the model-assisted calibration methods can improve the weight distributions relative to the model-free calibration, whose weight performance in small domains was expected to be unacceptable.

To examine empirically the weight performance we conducted a small simulation experiment with $K = 100$ SRSWOR samples from the synthetic population and computed the interquartile range (IQR) of calibrated weights for each method and each sample. Results are in Table 7.

Median IQR figures are smaller for all model-assisted methods relative to the reference estimator. Hájek methods outperform HT methods in all domain size groups, indicating more stable weight distribution than the HT methods. Best behaviour in small domains was attained with model-assisted Hájek type MC method. In both

model types, the two-level HC method competes well and stabilizes significantly the HC method.

TABLE 7. - *Median interquartile range (IQR) of Horvitz-Thompson and Hájek type calibrated weights relative (%) to the IQR of the HT type model-free calibration estimator in minor domains (synthetic population)*

Method	**Horvit-Thompson type weights**			**Hájek type weights**		
	Expected domain sample size (in parentheses)			Expected domain sample size (in parentheses)		
	Minor (12)	Medium (40)	Major (122)	Minor (12)	Medium (40)	Major (122)
Model-free estimator						
MFC	100	51	30	79	43	26
Model-assisted estimators						
MC	61	35	21	43	28	14
HC	78	43	26	60	36	21
HC2	61	36	21	47	28	16

The variation of weights is depicted graphically for Horvitz-Thompson and Hájek type methods in Figures 1 and 2. The HC2 weights outside domain d are omitted from the graphs for clarity. Our experiences have shown that outside domain d the weights will be close to zero in absolute value.

The HT weights in Figure 1 show very large variation for model-free calibration when domain sample sizes are small, and the variation stabilizes when domain sample sizes increase. Model-assisted calibration clearly improves weight behaviour relative to MFC, and there were fewer negative weights. Hybrid calibration also improves the weight distribution, but not substantially. The HT type two-level calibration behaves almost as well as model-assisted calibration. Surprisingly, negative weights have completely disappeared for Hájek type model-assisted calibration and two-level hybrid calibration in Figure 2, and their weight distributions are much more stable than weights for the HT type methods.

5.3 *Is HC2 design biased?*

The new two-level hybrid calibration HC2 estimator introduced in Section 3 is of indirect type by the construction principle. In this section we examine empirically the bias properties of Horvitz-Thompson type HC2 for domain total by simulation experiments. The experiments were planned to bring about large bias in a design biased estimator but small bias in a (nearly) design unbiased estimator. In our synthetic population (Section 5.1), the data generating model was substantially altered

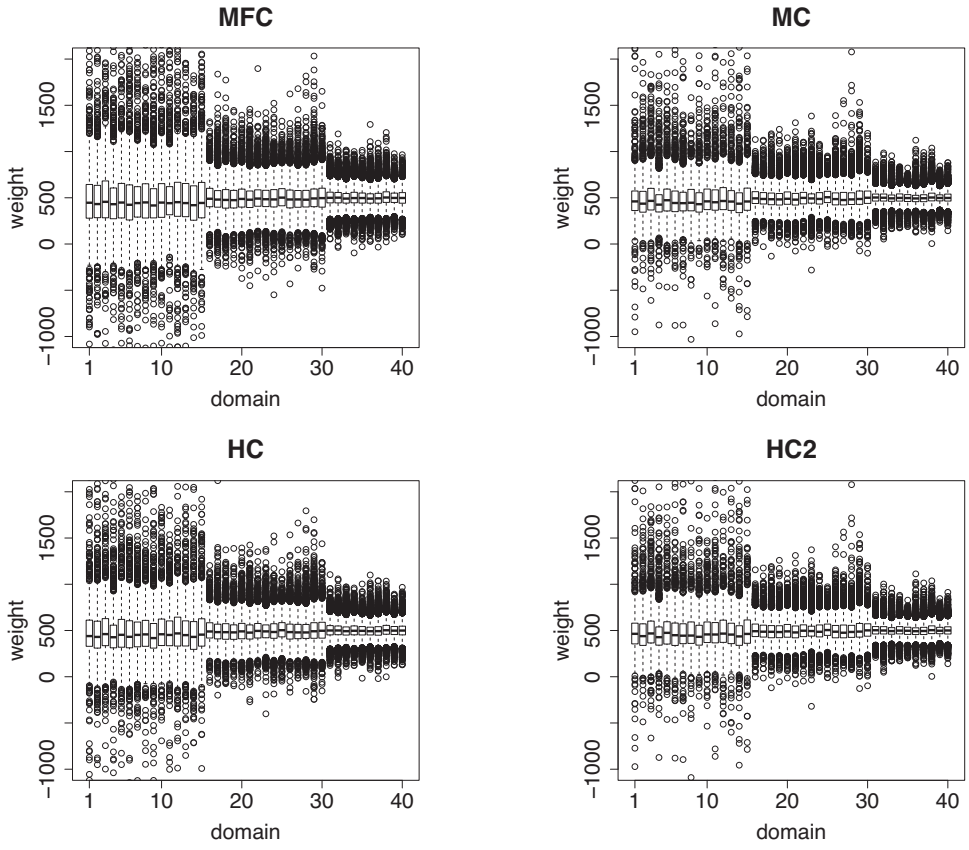

FIGURE 1. - *Distribution of weights in domains ordered by size for HT type model-free (MFC), model-assisted (MC), hybrid (HC) and two-level hybrid (HC2) calibration estimators (Synthetic population,* $K = 100$ *simulated SRSWOR samples of* $n = 2000$*)*

in a single small domain, named ''the outlying domain''. The altered model in this domain had distinctly larger intercept and regression coefficients than the models for the other domains. The model fitted to each sample will therefore deviate substantially from the true model in the outlying domain.

A standard unit-level EBLUP estimator commonly used in model-based small area estimation practice (e.g. Rao and Molina, 2015; Tzavidis, Zhang, Luna, Schmid and Rojas-Perilla, 2018) was selected as the model-based reference estimator. Under wrong model specification, the design bias of EBLUP estimator can be substantial (e.g. Lehtonen *et al.*, 2003). We included the Horvitz-Thompson type MFC, MC and HC estimators as design-based reference estimators. These design-

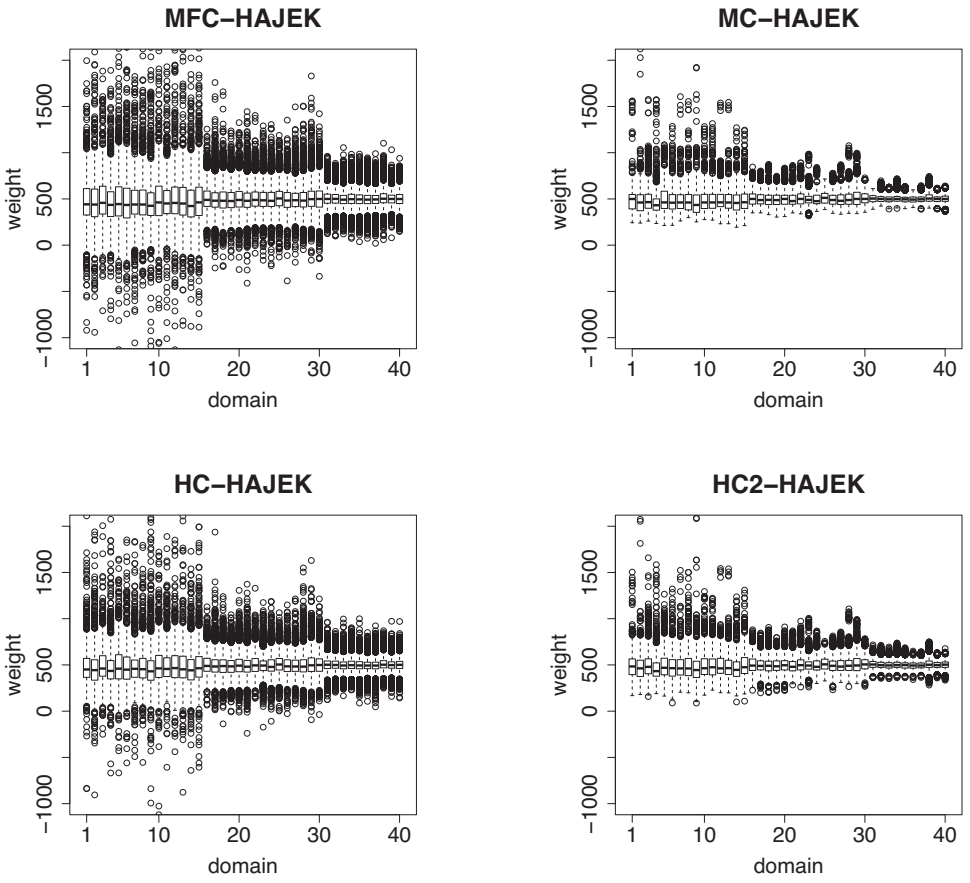

FIGURE 2. - *Distribution of weights in domains ordered by size for Hájek type model-free (MFC), model-assisted (MC), hybrid (HC) and two-level hybrid (HC2) calibration estimators (Synthetic population,* $K = 100$ *simulated SRSWOR samples of* $n = 2000$*)*

based estimators are by definition nearly design unbiased even under model misspecification. For a nearly design unbiased estimator, the design bias is, under mild conditions, an asymptotically insignificant contribution to the estimator's MSE (Särndal, 2007, p. 99).

We show that the bias ratio (empirical bias divided by empirical root MSE) is close to zero for HC2, verifying the near design unbiasedness property. We also verify that the empirical bias of HC2 is negligible and comparable to the biases of the nearly design unbiased reference estimators. In addition, the accuracy of HC2 is associated with the configuration of the auxiliary data supplied to the estimation procedure. It appears beneficial to incorporate the most powerful auxiliary data in the domain level model.

In the population, the auxiliary variables x_1 and x_2 were strongly correlated with the target y-variable, whereas the correlation of x_3 with y was close to zero. All three x-variables were incorporated in the estimation procedures for EBLUP, MFC and MC. A linear mixed model

$$y_k = \beta_0 + u_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \varepsilon_k, k \in U_d, d = 1, \ldots, D,$$

with domain-specific random intercepts was specified for the model-assisted estimators and the EBLUP estimator. Note that the model did not account for the anomalous structure of the outlying domain. For hybrid calibration estimators, the auxiliary information was supplied with two data vectors, the calibration x-vector and the model x-vector. For a fair comparison between the methods, the model x-vector and calibration x-vectors were defined as separate vectors, with no common x-variables. To exercise control over the impact of these vectors in the experiment, two different designs were attached to HC and HC2, as described in the setup below.

TABLE 8. - *Designs for model and calibration vectors*

	Model x-vector	Calibration x-vector
Design 1	$\mathbf{x}_{Mk} = (1, x_{1k}, x_{2k})'$	$\mathbf{x}_{Ck} = x_{3k}$
Design 2	$\mathbf{x}_{Mk} = (1, x_{3k})'$	$\mathbf{x}_{Ck} = (x_{1k}, x_{2k})'$

Design 1 represents the ''strong model x-vector, weak calibration x-vector'' case, because the most significant x-variables are in the model x-vector and the insignificant x-variable occupies the calibration x-vector. Design 2 represents the opposite case. The set-up involves calibration z-vectors summarized in Table 9. A single z-vector is attached to the HC estimator. The two-level HC estimator involves separate z-vectors. The first vector defines the model x-vector at the domain level (MC part), whereas the second one is for the calibration x-vector defined at the higher (regional) level (MFC part).

TABLE 9. - *Calibration z-vectors for the target and reference estimators*

	Design 1	Design 2
HC	$\mathbf{z}_k = (1, \hat{y}_k, x_{3k})'$ $\hat{y}_k = \hat{\beta}_0 + \hat{\boldsymbol{u}}_{0d} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k}$	$\mathbf{z}_k = (1, \hat{y}_k, x_{1k}, x_{2k})'$ $\hat{y}_k = \hat{\beta}_0 + \hat{\boldsymbol{u}}_{0d} + \hat{\beta}_3 x_{3k}$
HC2	$\mathbf{z}_k^{(1)} = (1, \hat{y}_k)'$ $\hat{y}_k = \hat{\beta}_0 + \hat{\boldsymbol{u}}_{0d} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k}$ $\mathbf{z}_k^{(2)} = x_{3k}$	$\mathbf{z}_k^{(1)} = (1, \hat{y}_k)'$ $\hat{y}_k = \hat{\beta}_0 + \hat{\boldsymbol{u}}_{0d} + \hat{\beta}_3 x_{3k}$ $\mathbf{z}_k^{(2)} = (x_{1k}, x_{2k})'$
MFC	$\mathbf{z}_k = (1, x_{1k}, x_{2k}, x_{3k})'$	
MC	$\mathbf{z}_k = (1, \hat{y}_k)', \hat{y}_k = \hat{\beta}_0 + \hat{\boldsymbol{u}}_{0d} + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \hat{\beta}_3 x_{3k}$	

In order to assess the near design unbiasedness (Särndal, 2007, p. 105) of the estimators of domain totals, we computed the relative empirical bias (43) of the estimators and absolute differences $RB(\hat{t}_d) - RB(\hat{t}_{dHT})$ of the relative biases of each calibration estimator $\hat{t}_d$ and the HT estimator. Means of these statistics were then computed over each domain sample size class. Small relative bias of an estimator of domain total and small absolute difference between empirical biases of a nearly design unbiased calibration estimator and the design unbiased HT estimator would indicate acceptable bias for the calibration estimator. The results under Design 1 are presented in Table 10. It is evident that the empirical design bias of HC2 does not differ from the other calibration estimators on average.

TABLE 10. - *Mean relative bias (%) of HT and calibration estimators of domain totals and mean absolute differences of relative biases (%) of calibration estimators vs. HT estimator in three domain sample size classes under Design 1 for the modified population*

Method	*Outlying minor domain*	*Expected domain sample size*		
		Minor (12)	Medium (40)	Major (122)
Mean relative bias (%)				
HT	−0.091	0.098	−0.003	−0.015
MFC	0.003	−0.031	0.002	−0.000
MC	−0.003	0.020	0.000	−0.000
HC	0.010	0.009	−0.000	−0.000
HC2	−0.002	0.014	0.000	−0.000
Mean absolute differences of relative biases (%)				
MFC vs. HT		0.288	0.136	0.064
MC vs. HT		0.246	0.139	0.064
HC vs. HT		0.263	0.137	0.064
HC2 vs. HT		0.248	0.139	0.064

We further examined the bias ratio (Särndal *et al.* 1992, p. 164-166) of the calibration estimators and the reference estimator. Bias ratio is obtained by dividing the empirical bias by empirical root mean squared error RMSE of an estimator of domain total, given by

$$BR(\hat{t}_d) = \frac{B(\hat{t}_d)}{RMSE(\hat{t}_d)} = \frac{\frac{1}{K}\sum_{j=1}^{K}\left(\hat{t}_d(s_j) - t_d\right)}{\sqrt{\frac{1}{K}\sum_{j=1}^{K}\left(\hat{t}_d(s_j) - t_d\right)^2}}, \qquad k \in U_d, d = 1, \ldots, D.$$

For a near design unbiased estimator the bias ratio should tend to zero as $n_d^{-1/2}$.

Results on bias ratio are in Table 11. The mean empirical bias ratios of the design-based calibration estimators, including HC2, are small and comparable on average. Mean bias ratio of the reference model-based EBLUP estimator in minor domains is ten times larger than the design-based counterparts and clearly larger in other domains. The results give support to the near design unbiasedness of the indirect HC2 estimator.

TABLE 11. - *Mean bias ratio (%) of HT and calibration estimators of domain totals in three domain sample size classes under Design 1 for the modified population*

Method	*Expected domain sample size* (in parentheses)		
	Minor (12)	Medium (40)	Major (122)
HT	0.87	0.89	0.79
MFC	0.75	0.80	0.77
MC	0.82	0.75	0.71
HC	0.75	0.76	0.71
HC2	0.76	0.75	0.73
EBLUP	7.65	1.45	1.26

The strength of auxiliary information in the domain-level and higher level parts of HC2 did not cause significant bias differences between the two designs. For accuracy, however, the strength of auxiliary data does matter. HC2 under Design 1 clearly outperforms HC2 under Design 2 in accuracy, in all domain size classes (Table 12). In the outlying domain, the accuracy was noticeably better under Design 1 than Design 2. For Design 1, the powerful auxiliary variables were inserted in the

TABLE 12. - *Mean RMSE and RRMSE (%) of two-level hybrid calibration estimator of domain totals in three domain sample size classes under designs 1 and 2 for the modified population*

	Outlying minor domain	*Expected domain sample size*		
		Minor (12)	Medium (40)	Major (122)
Mean RMSE				
Design 1	12956	10013	16264	28013
Design 2	19466	12083	19133	32882
Mean RRMSE (%)				
Design 1	1.66	4.03	1.67	0.79
Design 2	2.49	4.77	1.96	0.93

model defined at the domain level, and the weak auxiliary data were in the MFC part at the higher, regional level. In Design 2, the HC2 estimator incorporated domain level information about the weak auxiliary variable and the strong auxiliary variables were included at the regional level only. This implies that the model incorporated in HC2 should contain at least some good auxiliary variables that explain differences between domains. Including more auxiliary information at regional level in HC2 does not necessarily compensate for the lack of sufficient domain level information.

6. SIMULATION EXPERIMENTS WITH REAL DATA

6.1 *Accuracy comparison*

We compared empirically the bias and accuracy of the various calibration estimators by design-based Monte Carlo experiments applied to a real population constructed from the income-related statistical registers of Statistics Finland. The binary poverty indicator shows when a person's equivalized income U_k is smaller than or equal to the poverty threshold, 60% of the median equivalized income M in the population. The indicator for sample person k is defined as $y_k = I\{U_k \leq 0.6\hat{M}\}$, where $y_k = 1$ if a person is in poverty and 0 otherwise. The quantity $0.6\hat{M}$ is the estimated poverty threshold, where $\hat{M}$ was estimated by HT from the estimated distribution function of equivalized income in the population (Lehtonen and Veijanen, 2012). The binary poverty indicator was used as the target variable for the study. An adult population of about 800000 persons was constructed, containing 36 LAU level 1 regions in Western Finland, organized hierarchically within seven NUTS level 3 regions. In addition to the equalized income variable, the population contained three auxiliary variables: two-category gender, four-category age and two-category labour force status. We created indicator variables for classes of each qualitative variable; one indicator for sex and labour force status and three indicators for age. The complete auxiliary x-vector for $k \in U$ is $\mathbf{x}_k = (x_{0k}, x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k}, x_{6k})'$, where $x_{0k} = 1$. The covariates indicated modest explanatory power: in logistic models, the complete x-data explained about 15% of the variation of the target variable. As domains of interest U_d we used the $D = 36$ LAU-1 domains. The seven NUTS-3 regions constitute the higher-level $U_{r(d)}$ regions. Overall poverty rate in population was 14.3%. In the regions, lowest rate was 9.9% and highest was 22.4%.

In the simulations, $K = 1000$ samples of $n = 2000$ units were drawn with SRSWOR from the unit-level population. Because domain sample sizes were random, some samples were ignored because of too small realized sample size. In the class of small domains, 0.7% of the HC2 estimates, 1.1% of HC estimates and 3.3% of MFC estimates were rejected. No MC estimates were rejected. The rejection of estimates in some methods does not change the overall picture of the properties of the methods.

The parameter of interest was domain poverty rate (2). Design bias and accuracy of domain poverty estimators $\hat{p}_d$ were measured by absolute relative bias (ARB) and

relative root mean squared error (RRMSE) as given in (41) and (42). Poverty rate (2) for domain d was estimated by HT and Hájek type calibration estimators $\hat{p}_d^{HT} = \hat{t}_d^{HT}/N_d$ and $\hat{p}_d^{HA} = \hat{t}_d^{HA}/N_d$, $d = 1, \ldots, D$, obtained by (16), (24), (32) and (37) for HT and (17), (25), (33) and (38) for HA. The calibration vectors are collected in Table 13.

TABLE 13. - *Calibration vectors for the Horvitz-Thompson (HT) and Hájek (HA) type estimators*

MFC	HT calibration vector $\mathbf{z}_i = (1, \mathbf{x}_{Ci}')'$ HA calibration vector $\mathbf{z}_i = \mathbf{x}_{Ci}$, $i \in U_d$ $\mathbf{x}_{Ci} = (x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i})'$
MC	Model x-vector $\mathbf{x}_{Mi} = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i})'$ HT calibration vector $\mathbf{z}_i = (1, \hat{y}_i)'$ HA calibration vector $\mathbf{z}_i = \hat{y}_i, i \in U_d$
HC	Model x-vector $\mathbf{x}_{Mi} = (1, x_{1i}, x_{2i})'$ for LF status Calibration x-vector $\mathbf{x}_{Ci} = (x_{3i}, x_{4i}, x_{5i}, x_{6i})'$ for gender and age group HT calibration vector $\mathbf{z}_i = (1, \hat{y}_i, \mathbf{x}_{Ci}')'$ HA calibration vector $\mathbf{z}_i = (\hat{y}_i, \mathbf{x}_{Ci}')'$, $i \in U_d$
HC2	Model x-vector $\mathbf{x}_{Mi}$ and calibration x-vector $\mathbf{x}_{Ci}$ as in HC Calibration vectors HT Level 1: $\mathbf{z}_i^{(1)} = (x_{0i}^{(1)}, \hat{y}_i^{(1)})', i \in U_r(d)$ HA Level 1: $\mathbf{z}_i^{(1)} = \hat{y}_i^{(1)}, i \in U_{r(d)}$ $x_{0i}^{(1)} = 1, \hat{y}_i^{(1)} = \hat{y}_i, i \in U_d; x_{0i}^{(1)} = 0, \hat{y}_i^{(1)} = 0, i \in U_{r(d)} \setminus U_d$ HT and HA Level 2: $\mathbf{z}_i^{(2)} = \mathbf{x}_{Ci}, i \in U_{r(d)}$
Predictions for MC, HC, HC2	$\hat{y}_i = \exp\left(\mathbf{x}_{Mi}'\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_{0d}\right) / \left(1 + \exp\left(\mathbf{x}_{Mi}'\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_{0d}\right)\right), i \in U_d$

Our logistic mixed models (4) contained regional random intercepts associated with LAU-1 domains. We computed median ARB and RRMSE over three domain size classes defined by expected domain sample size. Results are presented in Table 14.

Median bias figures in the upper part of the table provide empirical support to the near unbiasedness of the calibration estimators, even for the group of small domains, and there were no significant differences between the methods. Results on accuracy are in the lower part of the table. Over all domains, model-assisted calibration outperforms direct model-free calibration in accuracy, for both HT and Hájek type methods. This holds for all domain sample size classes and is best visible in the minor and medium-sized groups. In model-assisted methods, both method families show similar accuracy in the major domains, Hájek type methods are slightly more accurate in the small domains group. In both families, model-assisted calibration indicates best accuracy in this size group. Hybrid calibration suffers from instability, which is again reduced by the two-level HC method. The logistic mixed model clearly tends

to improve accuracy over the direct method, whose implicit assisting model is a linear fixed-effects model fitted separately in each domain. We reached similar conclusions as with the synthetic population.

TABLE 14. - *Median absolute relative bias ARB (%) and root mean squared error RRMSE (%) of Horvitz-Thompson and Hájek type calibration estimators for domain totals in three domain sample size classes (real population)*

Method	Horvitz-Thompson type estimators				Hájek type estimators			
	Expected domain sample size			All	Expected domain sample size			All
	Minor < 25	Medium $25-50$	Major > 50		Minor < 25	Medium $25-50$	Major > 50	
Median ARB (%)								
Model-free estimator								
MFC	1.96	1.60	0.64	1.19	1.98	1.49	0.64	1.15
Model-assisted estimators								
MC	2.12	1.11	0.58	1.11	1.57	0.88	0.71	1.05
HC	2.38	1.28	0.68	1.17	2.91	1.32	0.78	1.32
HC2	1.49	0.94	0.76	0.92	1.49	0.83	0.77	0.83
Median RRMSE (%)								
Model-free estimator								
MFC	70.8	48.7	30.9	48.7	64.6	47.7	30.6	47.7
Model-assisted estimators								
MC	54.6	44.0	29.9	44.0	53.9	43.6	30.2	43.6
HC	69.5	47.1	30.6	47.1	64.1	47.5	30.9	47.5
HC2	55.2	44.2	30.0	44.2	54.2	44.1	30.4	44.1

6.2 *Distribution of weights*

We conducted a small simulation experiment similar to Section 5.2 to examine the weight distributions of the calibration methods. A total of $K = 100$ SRSWOR samples of 2000 units were drawn from the real population. The variation of weights was illustrated by plotting the weights against expected domain sample size, for the four competing calibration methods in both HT and Hájek type families. Results are in Figures 3 and 4.

Huge variation of weights is obtained in Figure 3 for the HT type MFC and HC

methods. Weights stabilize slowly only when domain sample sizes increase. The assisting logistic model makes the HC weights behave slightly better than the MFC weights. The strength of the model is best visible in model-assisted calibration. Weights are stable and extreme weights and negative weights are rare. The MC and two-level hybrid calibration methods provide the most stable weight performance. Weights of Hájek type MFC and HC perform similarly as the HT counterparts, as can be seen in Figure 4. Dramatic stabilization in weight distributions takes place when turning to the MC and HC2 methods. The model-assisted calibration and two-level hybrid calibration methods appear to be effective weight stabilization methods, if the assisting model is strong enough. The bonus in HC2 is the attained coherence of the selected auxiliary variables at a higher regional level.

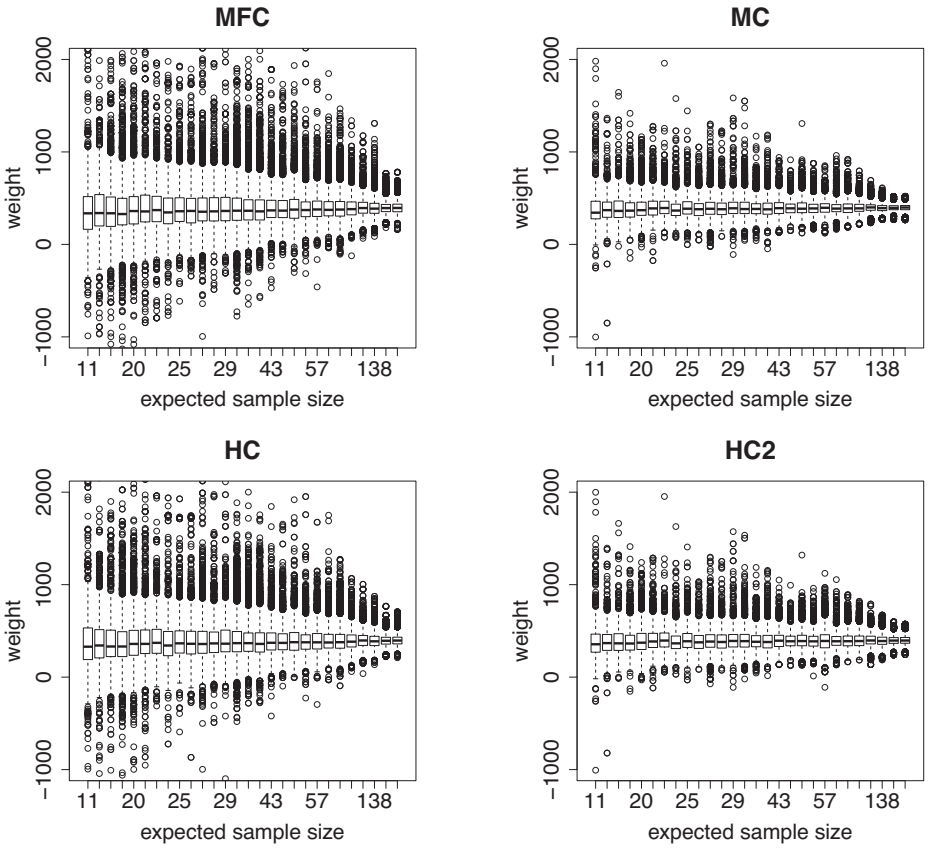

FIGURE 3. - *Distribution of weights in domains ordered by size for HT type model-free (MFC), model-assisted (MC), hybrid (HC) and two-level hybrid (HC2) calibration estimators (Real population, K = 100 simulated SRSWOR samples of n = 2000)*

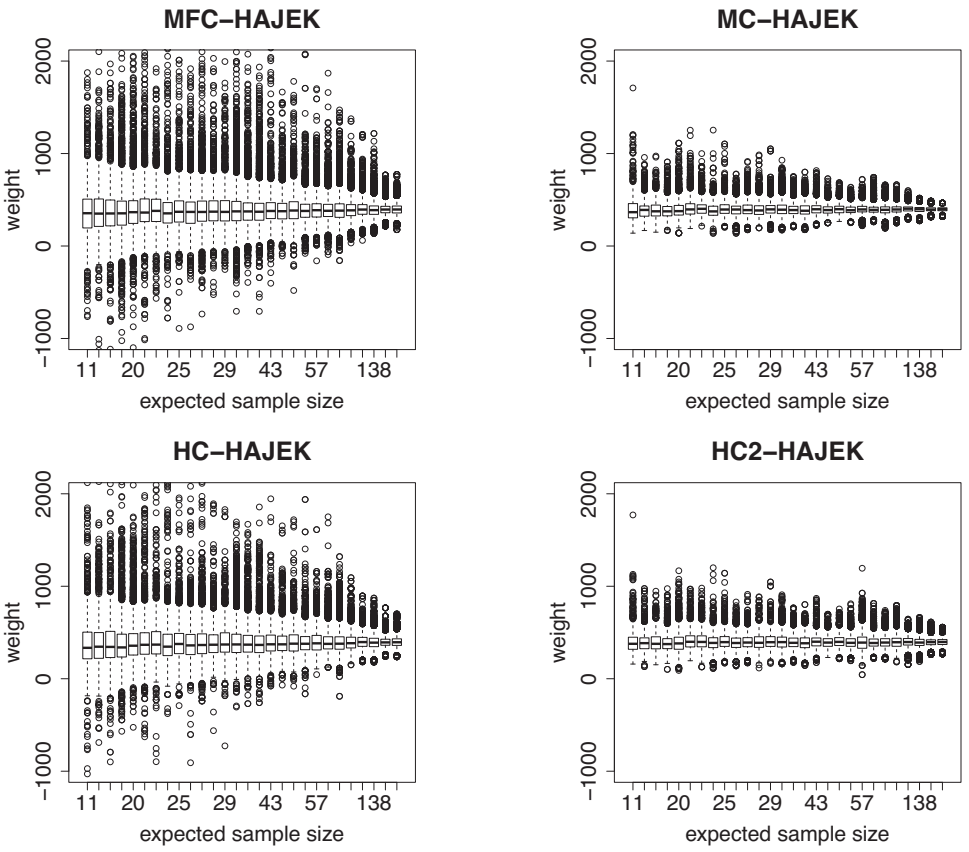

FIGURE 4. - *Distribution of weights in domains ordered by size for Hájek type model-free (MFC), model-assisted (MC), hybrid (HC) and two-level hybrid (HC2) calibration estimators (Real population, K = 100 simulated SRSWOR samples of n = 2000)*

7. SUMMARY AND DISCUSSION

The key feature of the design-based model-assisted calibration methods introduced here is their ability to incorporate flexible modelling into the calibration procedure. This property extends the calibration approach to the estimation for population subgroups or domains whose realized sample sizes are small, where the direct estimates from the classical model-free calibration become unstable. Model-assisted calibration also extends the classical calibration beyond linear modelling of continuous variables towards other variable types met in practice, and offers an option to include random effects in the assisting models to account for the possible heterogeneity over the domains. We adopted linear and logistic mixed models as assisting models in

deriving the calibration estimators for small domain estimation of domain totals and proportions. The models together with the available unit-level auxiliary variables were used to produce predicted values for population units to be incorporated in the model-assisted calibration procedures.

Calibration estimators were constructed by three ways: (a) with the original auxiliary variables, (b) the pseudo auxiliary data i.e. predictions, or (c) a combination of (a) and (b). The classical MFC relies to the first option. Of the model-assisted methods, basic MC uses solely the predictions. The two hybrid methods, HC and HC2, use (c). As statistical infrastructures differ, each of these methods has its role in calibration estimation.

In the MC method, the weighted sample sums of predictions in domains were calibrated to the domain sums of the predictions in the population. Thus, the coherence property of classical model-free calibration to reproduce the published statistics of auxiliary variables was lost. In hybrid calibration, both the predictions and the original auxiliary variables were used, creating a HC estimator. In HC, there is a model-free calibration part and a model-calibration part. HC thus allows the coherence property for selected auxiliaries at the original domain level, and still retains the accuracy improvement property of MC. To avoid instability problems inherent in the MFC part of HC in small domains, we developed a new HC2 method. In HC2, calibration for the original auxiliary variables was applied at a higher regional level, instead of the domain level as in HC, and calibration for the model predictions was conducted at the domain level. To demonstrate the flexibility of hybrid calibration methods we applied both the separate and overlapping auxiliary variable configurations in defining the share of the auxiliary variables between the MFC and MC parts of the calibration procedure.

Calibration estimators were derived under two approaches, the HT approach and the HA approach. The HT type calibration estimators of total incorporated calibrating the sample sums of weights to the known domain sizes, whereas this was not the case for Hájek type calibration estimators. The Hájek type estimators were simple generalizations of the classical Hájek estimator of population total under PPS sampling (Hájek, 1971). Domain proportion estimators were obtained by dividing a HT or HA estimator by the known domain size in population.

We examined the design bias and accuracy of the calibration estimators empirically by design-based simulation experiments using large artificially generated and real populations. Our main interest was in small domains, but we presented results for three domain sample size classes in order to give a clear picture of the properties of the estimators for domains of different sample sizes. In addition, we were interested in the stability of the calibrated weights of each method. The distributions of weights were examined by small simulation experiments and were illustrated with graphical presentations.

All the design-based estimators developed in this study appeared nearly design unbiased. The indirect two-level hybrid calibration estimator, in particular, had small design bias. Conclusions on bias were similar for both the Horvitz-Thompson and Hájek type estimators, and for the synthetic and real populations.

The calibration estimators showed nearly identical accuracy for medium-sized and large domains, in both populations. The basic model-assisted MC estimator was the most accurate method in small domains, and the distribution of weights were stable. Weights from the Hájek type MC method were surprisingly stable in the real population in particular. For MC, extreme and negative weights were absent or rare even in small domains, contrary to the model-free calibration that suffered from these problems. In hybrid calibration, the price to be paid for attaining the coherence property for some auxiliary variables was declined accuracy relative to the MC method in small domains, and the weights were unstable. The two-level HC method, where the coherence property was assigned to a higher regional level, improved significantly the accuracy over the HC method. The weight distributions were stable and extreme and negative weights were rare, similarly as in the basic model-assisted MC method. In the group of small domains in the synthetic population, the Horvitz-Thompson type model-assisted calibration estimators were slightly more accurate than the Hájek type estimators; the situation was opposite for the real population, but the differences were minor.

The two-level hybrid calibration behaved well in the situations considered in this study and may offer a safe compromise if coherence is required for some auxiliary variables and the option for efficiency improvement via modelling is desired. If coherence is not an issue, both variants of hybrid calibration provide alternatives for hierarchically structured populations if information is available at different levels of the population, including unit level values for some auxiliary variables and aggregates at the various higher levels for other variables, such as auxiliary variables and even the target y-variable. The flexibility comes from the fact that the model-free part of the hybrid calibration procedure does not involve auxiliary x-variable values at the unit level, whereas unit-level variables are assumed for the model-assisted MC part.

For precision assessment of MC, HC and HC2 estimators, we developed simple versions of a variance estimator presented by Kott (2009) and examined their design-based properties by simulation experiments. For MC, design bias and accuracy properties of the MSE estimator were reasonably good. MSE estimator of HC2 was comparable to MC in accuracy but suffered from negative design bias in small domains. For HC, MSE estimation was unreliable in the small domains.

The basic model-assisted calibration and two-level hybrid calibration methods provided effective approaches for weight stabilization in small domain estimation in order to avoid too large and negative weights. As assisting models we used members of the generalized linear mixed models family. Nonparametric methods sometimes used in model-based small area estimation (e.g. Ranalli, Breidt and Opsomer, 2016) may provide an alternative for weight stabilization in model-assisted small domain estimation. This will remain a subject of future research.

The linear calibration method with a chi-square type distance function used in our studies may involve extreme weights and negative weights, which are often considered unfeasible. This was observed also in our studies for the classical model-free calibration in particular. Many techniques have been proposed in the literature to restrict the variation of weights and for weight trimming and smoothing (e.g. Deville

and Särndal, 1992, p. 378; Chen, Sitter and Wu, 2002; Park and Fuller, 2005; Beaumont and Bocci, 2008; Guggemos and Tillè, 2010; Kim, 2010; Wu and Lu, 2016). Alternative (asymptotically equivalent) distance functions are possible in order to avoid unfeasible weights (Deville and Särndal, 1992, Section 2).

We did not incorporate weight restriction techniques or alternative distance measures in the calibration procedures, although they are popular in the calibration practice. The choice of the limits is more or less arbitrary. It is not clear how these techniques behave in small domain estimation, and more research is needed in this area. We wanted to examine the methods in a pure framework without disturbing subjective elements.

In survey practice, empty domains (domains with zero sample size) are possible when working with ad hoc domain structures. In this study we assumed nonzero sample size for every domain of interest, because a well-controlled framework without unnecessary complexities is optimal in assessing the relative properties (design bias, accuracy) of new methods against competitors. The framework of this study does not readily extend to very small or zero domain sample sizes. Our design-based methods for small domain estimation appeared to work well at least for realized sample sizes of about ten elements. Extension to the empty domains cases needs further research.

ACKNOWLEDGEMENTS

We thank Emilia Rocco and Pier Francesco Perri for their cooperation and a referee for suggesting significant improvements to the manuscript.

REFERENCES

Basu D. (1971). An essay on the logical foundations of survey sampling, part I (with discussion). In V.P. Godambe and D.A. Sprott (Eds.), *Foundations of Statistical Inference* (pp. 203-242). Holt, Rinehart and Winston, Toronto.

Beaumont J.-F., Bocci C. (2008). Another look at ridge calibration. *Metron*, **LXVI**, 5-20.

Breidt F.J., Opsomer J.D. (2009). Nonparametric and semiparametric estimation in complex surveys, theory, methods and inference. In C.R. Rao and D. Pfeffermann (Eds.) *Handbook of Statistics, Vol. 29B. Sample Surveys: Inference and Analysis*. (pp. 103-119). Elsevier, Amsterdam.

Burgard J.P., Dörr P. (2018). Survey-weighted generalized linear mixed models. *Research Papers in Economics 2018-01*. University of Trier, Department of Economics.

Burgard J.P., Münnich R., Rupp M. (2019). A generalized calibration approach ensuring coherent estimates with small area constraints. *Universität Trier, Research Papers in Economics*,

No. **10/19**. Accessed in 24 October 2019 at: https://www.uni-trier.de/fileadmin/fb4/prof/VWL/EWF/Research-Papers/2019-10.pdf

Canty A.J., Davison A.C. (1999). Resampling-based variance estimation for Labour Force Surveys. *The Statistician*, **48**, 379-391.

Chandra H., Chambers R. (2011). Small area estimation under transformation to linearity. *Survey Methodology*, **37**, 39-51.

Chen J., Sitter R.R., Wu C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, **89**, 230-237.

Chen J.K.T., Valliant R.L., Elliott M.R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, **44**, 117-144.

Dagdoug M., Goga C., Haziza D. (2020). Model-assisted estimation through random forests in finite population sampling. https://arxiv.org/pdf/2002.09736.pdf

Deville J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, **25**, 193-203.

Deville J.-C., Särndal C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

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

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

Fabrizi E., Salvati N., Pratesi M., Tzavidis N. (2014). Outlier robust model-assisted small area estimation. *Biometrical Journal*, **56**, 157-175.

Federal Committee on Statistical Methodology (1993). Indirect Estimators in Federal Programs. U.S. *Office of Management and Budget, Statistical Policy Working Paper*, **21**. Accessed in 24 October 2019 at: https://nces.ed.gov/FCSM/pdf/spwp21.pdf

Guggemos F., Tillè Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference*, **140**, 3199-3212.

Hájek J. (1971). Comment on ''An essay on the logical foundations of survey sampling'' by Basu, D. In V.P. Godambe and D.A. Sprott (Eds.) *Foundations of Statistical Inference* (p. 236). Holt, Rinehart and Winston, Toronto.

Hidiroglou M.A., Estevao V.M. (2016). A comparison of small area and calibration estimators via simulation. *Joint Issue of Statistics in Transition and Survey Methodology*, **17**, 133-154.

Hidiroglou M.A., Patak Z. (2004). Domain estimation using linear regression. *Survey Methodology*, **30**, 67-78.

Horvitz D.G., Thompson D.J. (1952). A generalization of sampling with-out replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

Huang E.T., Fuller W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305. Washington, DC.

Kim J.K. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology*, **36**, 145-155.

Kim J.K., Park M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, **78**, 21-39.

Kott P.S. (2009). Calibration weighting: combining probability samples and linear prediction models. In C.R. Rao and D. Pfeffermann (Eds.), *Handbook of Statistics, Vol. 29B. Sample Surveys: Inference and Analysis* (pp. 55-82). Elsevier, Amsterdam.

Lehtonen R., Veijanen A. (2009). Design-based methods of estimation for domains and small areas. In C.R. Rao and D. Pfeffermann (Eds.), *Handbook of Statistics, Vol. 29B. Sample Surveys: Inference and Analysis* (pp. 219-249). Elsevier, Amsterdam.

Lehtonen R., Veijanen A. (2012). Small area poverty estimation by model calibration. *Journal of the Indian Society of Agricultural Statistics*, **66**, 125-133.

Lehtonen R., Veijanen A. (2016). Design-based methods to small area estimation and calibration approach. In M. Pratesi (Ed.), *Analysis of Poverty Data by Small Area Estimation* (pp. 109-127). Wiley, Chichester.

Lehtonen R., Särndal C.-E., Veijanen A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, **29**, 33-44.

Lehtonen R., Särndal C.-E., Veijanen A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, **7**, 649-673.

McConville K.S., Breidt F.J., Lee T.C.M., Moisen G.G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, **5**, 131-158.

Montanari G.E., Ranalli M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, **100**, 1429-1442.

Montanari G.E., Ranalli M.G. (2009). Multiple and ridge model calibration. *Proceedings of Workshop on Calibration and Estimation in Surveys 2009*. Statistics Canada. Accessed in 24 October 2019 at: http://www.stat.unipg.it/~giovanna/papers/ranalliWCES.pdf

Morales D., Rueda M., Esteban D. (2018). Model-assisted estimation of small area poverty measures: an application within the Valencia region in Spain. *Social Indicators Research*, **138**, 873-900.

Park M., Fuller W.A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology*, **31**, 85-93.

Ranalli M.G., Breidt J.F., Opsomer J.D. (2016). Nonparametric regression methods for small area estimation. In M. Pratesi (Ed.), *Analysis of Poverty Data by Small Area Estimation* (pp. 188-204). Wiley, Chichester.

Rao J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhya, Series A.*, **28**, 47-60.

Rao J.N.K., Molina I. (2015). *Small Area Estimation*, 2nd ed. Wiley, New York.

Rueda M., Arcos A., Molina D., Trujillo M. (2018). Model-assisted and model-calibrated estimation for class frequencies with ordinal outcomes. Manuscript. Accessed in 24 October 2019 at: https://ine.pt/revstat/pdf/MODEL-ASSISTEDANDMODEL-CALIBRATEDESTIMATION.pdf

Rueda M., Sànchez-Borrego I., Arcos A., Martìnez S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, **71**, 33-44.

Särndal C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, **33**, 99-119.

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

Torabi M., Rao J.N.K. (2008). Small area estimation under a two-level model. *Survey Methodology*, **34**, 11-17.

Tzavidis N., Zhang L.-C., Luna A., Schmid T., Rojas-Perilla N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society, Series A*, **181**, 927-979.

Wang L., Wang S. (2011). Nonparametric additive model-assisted estimation for survey data. *Journal of Multivariate Analysis*, **102**, 1126-1140.

Wu C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, **90**, 937-951.

Wu C., Lu W.W. (2016). Calibration weighting methods for complex surveys. *International Statistical Review*, **84**, 79-98.

Wu C., Sitter R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185-193.