Journal of Computational and Applied Mathematics 446 (2024) 115876

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Calibration estimation of distribution function based on multidimensional scaling of auxiliary information

Sergio Martínez [a,*], María D. Illescas [b], María del Mar Rueda [c]

[a] Department of Mathematics. University of Almería, Ctra. Sacramento s/n, La Cañada de San Urbano, Almería, 04120, Spain
[b] Department of Economics and Business, University of Almería, Ctra. Sacramento s/n, La Cañada de San Urbano, Almería, 04120, Spain
[c] Department of Statistics and Operational Research. Institute of Mathematics (IMAG). University of Granada, Avenida de la Fuente Nueva S/N, Granada, 18071, Spain

ARTICLE INFO

MSC:
62D05

Keywords:
Sampling
Distribution function
Calibration
Multidimensional scaling

ABSTRACT

The distribution function is a functional parameter of great interest in many research areas, such as medicine or economics. Among other properties, it facilitates the estimation of parameters such as quantiles. Accordingly, techniques are needed to estimate this function efficiently.

Survey statisticians have access to large, high-dimension databases and use them to optimise the estimates obtained. One way to incorporate auxiliary information in the estimation stage is through the calibration method, which was initially designed to estimate totals and means and consists of adjusting new sample weights in order to reduce the variance of estimators. However, calibration techniques may be subject to over-calibration, i.e. the loss of efficiency when high-dimension auxiliary data sets are incorporated.

Although alternative approaches have been proposed, in which the calibration method incorporates auxiliary information in the estimation of the distribution function, these alternatives do not seek to incorporate qualitative auxiliary information, which must be introduced in the usual way through dummy variables. However, this workaround can greatly increase the dimension of the auxiliary information, producing either over-calibration or even incompatible calibration constraints.

In this article, we propose adapting the calibration method through multidimensional scaling, in order to incorporate quantitative and qualitative information, thus avoiding the negative consequences of over-calibration in the estimation of the distribution function.

1. Introduction

The estimation of the distribution function, a parameter that is non-linear and functional, is currently a significant topic in the context of sampling surveys. Among other reasons for its importance, in several cases the estimation of the distribution function is more helpful than that of the totals and the means [1], since this function allows us to obtain other parameters such as the reliability function [2], the Gini index [3,4], the Headcount index [5], the poverty incidence, the poverty gap and the poverty severity [6], as well as population quantiles [7–9], which are commonly addressed in research areas such as medicine [10], toxicology [11], edaphology [12] and economics [13].

Recent technological advances in automatic collection and storage capacity have increased the volume of information available and facilitated access to it [14]. Thus, survey statisticians, for example, can now consider an extensive range of variables linked to the population of interest, which can be incorporated as auxiliary information to improve the estimations obtained.

* Corresponding author.
E-mail addresses: spuertas@ual.es (S. Martínez), millescas@ual.es (M.D. Illescas), mrueda@ugr.es (M.d.M. Rueda).

https://doi.org/10.1016/j.cam.2024.115876
Received 3 September 2023; Received in revised form 15 December 2023
Available online 14 March 2024

In view of this ease of access to a significant volume of auxiliary information, together with the notable relevance of the distribution function, we believe it essential to derive indirect estimators of the distribution function that efficiently incorporate the available auxiliary information. One way to do so, in the estimation phase, is to use the calibration method, which was originally developed to estimate finite population totals or means [15]. Calibration provides a weighting system that satisfies a set of calibration restrictions and minimises a specified distance measure between the design weight system and the calibration weight system [15].

Some recent proposals have been made to adapt the calibration method to estimate the distribution function [16–21]. Among these, the approach described by [19] is especially useful, as it provides estimators that are true distribution functions under smooth requirements. Moreover, it offers computational simplicity. On the contrary, the asymptotic behaviour of the estimators discussed in [19] depends on the choice of an auxiliary vector [22,23] whose optimal selection can have a large dimension [24]. All of these previous studies assume that the auxiliary information available is quantitative, while none consider how we might incorporate auxiliary information of a qualitative nature. To the best of our knowledge, this issue has received very little attention in the context of calibration research [25].

Under present conditions, the volume of auxiliary information related to the target population may be high, and therefore it cannot be assumed that there does not exist a large set of qualitative variables included as auxiliary variables. To date, the incorporation of qualitative auxiliary variables in the estimation of the distribution function has been achieved by means of the corresponding dummy variables. However, this means of incorporating qualitative variables might further increase the dimension of the auxiliary information used in the calibration process.

Calibration is generally agreed to be a reliable method to incorporate auxiliary information and obtain new asymptotically unbiased estimators for several parameters [26,27], even in difficult sampling contexts such as when the sample has missing data [28], or when successive sampling is performed [29] or in the case of dual frame surveys [30]. However, the use of a set of high-dimension auxiliary variables in the calibration process can pose several major problems. Firstly, the calibration process might incorporate restrictions that are incompatible and/or unstable. Moreover, even if these restrictions are compatible, the estimators thus obtained may suffer from over-calibration when the dimension of the auxiliary information exceeds a certain threshold [31], which is quite plausible if a large set of qualitative variables are represented by dummy variables. Over-calibration can reduce the efficiency of the calibrated estimator [14], preventing it from making the best use of the large volume of auxiliary information made available.

In this context, various approaches have been proposed to reduce the dimension of the auxiliary information when calibrated estimation is used to determine totals and means [18,25,32–36]. According to [19], there are alternative means of reducing the dimension of the auxiliary information used in the calibration process when estimating the distribution function [37,38], although these methods assume that all the auxiliary variables are quantitative. Consequently, alternative approaches to dummy variables are necessary in order to incorporate qualitative information into the calibration process without considerably increasing the dimension of auxiliary information.

In order to overcome the limitations detected in previous approaches, and in line with [25] regarding the estimation of totals, the aim of the present study is to develop an alternative approach to calibration, based on multidimensional scaling (MDS), which makes it possible to incorporate all the auxiliary information available, both quantitative and qualitative, in estimating the distribution function. Under this new proposal, which integrates the approaches of [19,25], new calibration estimators are developed for the distribution function, seeking to incorporate both qualitative and quantitative information and thus avoid over-calibration.

2. Calibration estimators of the distribution function

Consider a finite population $U = \{1, \dots, N\}$ of size N and a given sampling design $p(\cdot)$ with inclusion probabilities of first and second order $\pi_k > 0$ and $\pi_{ki} > 0$ $k, i \in U$. Then, we denote by $d_k = \pi_k^{-1}$ the sampling design weight for unit $k \in U$. A sample $s = \{1, 2, \dots, n\}$ with fixed size n is selected according to the sampling design $p(\cdot)$ from the population U. If Y is the study variable we denote by y_k the value of the study variable for unit k that is only known for units included in s. We consider J auxiliary variables $X_1, \dots, X_J$ whose values are available for all population units (complete auxiliary information) and we denote by $\mathbf{x}_k' = (x_{1k}, \dots, x_{Jk})$ the vector with the J auxiliary variables at unit k. Our aim in this is to estimate the distribution function $F_y(t)$ for the study variable Y, given by:

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) \tag{1}$$

where

$$\Delta(t - y_k) = \begin{cases} 1 & \text{if } t \geq y_k \\ 0 & \text{if } t < y_k. \end{cases}$$

A well-known unbiased estimator for $F_y(t)$ is the Horvitz‚ÄìThompson estimator, which is defined by

$$\widehat{F}_{YHT}(t) = \frac{1}{N} \sum_{k \in s} d_k \Delta(t - y_k). \tag{2}$$

However, this estimator $\widehat{F}_{YHT}(t)$ does not take advantage of the auxiliary information provided by $\mathbf{x}_k$. On the other hand, the calibration method [15] could be used to incorporate this information in the estimation stage. Although it was originally developed in the estimation of totals or means, some studies have adapted it to estimate the distribution function [16–23].

Among these earlier proposals, in this study we focus on the approach described by [19] which provides estimators at modest computational cost. Under this approach, we assume that all variables included in the auxiliary vector $\mathbf{x}'_k$ are quantitative. Accordingly, we can define the following pseudo-variable:

$$g_k = \widehat{\beta}'\mathbf{x}_k \text{ for } k = 1, 2, \dots, N \tag{3}$$

$$\widehat{\beta} = \left(\sum_{k\in s} d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \cdot \sum_{k\in s} d_k \mathbf{x}_k y_k \tag{4}$$

Based on this pseudo-variable g, the calibration procedure replaces the design weight d_k in the Horvitz–Thompson estimator by a new calibration weight ω_k that meets the following calibration restrictions:

$$\frac{1}{N}\sum_{k\in s} \omega_k \Delta(t_j - g_k) = F_g(t_j) \quad j = 1, 2, \dots, P \tag{5}$$

and at the same time minimises the chi-square distance:

$$\Phi_s(\omega_k, d_k) = \sum_{k\in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{6}$$

where we assume that q_k are positive constants not related to d_k, $t_j \quad j = 1, 2, \dots, P$ are J points with $t_1 < t_2 < \cdots t_P$ and that $F_g(t_j)$ denotes the distribution function of g evaluated at the point t_j.

The resulting estimator is given by:

$$\widehat{F}_{yc}(t) = \widehat{F}_{YHT}(t) + \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g})\right)' \cdot \widehat{D}(\mathbf{t_g}) \tag{7}$$

where $\widehat{F}_{GHT}(\mathbf{t_g})$ denotes the Horvitz–Thompson estimator for $F_g(\mathbf{t_g})$ evaluated at $\mathbf{t_g} = (t_1, \dots, t_P)'$ and

$$\widehat{D}(\mathbf{t_g}) = T^{-1} \cdot \sum_{k\in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(t - y_k)$$

and where it is essential to assume that the matrix

$$T = \sum_{k\in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'$$

is nonsingular.

Among the advantages of $\widehat{F}_{yc}(t)$, [19] established that $\widehat{F}_{yc}(t)$ is a true distribution function if $q_k = c$ for all $k \in s$ and if t_P is large enough to guarantee $F_g(t_P) = 1$, so it can be used in the estimation of quantiles [8]. Additionally, $\widehat{F}_{yc}(t)$ is asymptotically unbiased and its asymptotic variance can be established by the following expression:

$$AV(\widehat{F}_{yc}(t)) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}(d_k E_k)(d_l E_l) \tag{8}$$

where $E_k = \Delta(t - y_k) - \Delta(\mathbf{t_g} - g_k)' \cdot D(\mathbf{t_g})$, with

$$D(\mathbf{t_g}) = \left(\sum_{k\in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'\right)^{-1} \cdot \left(\sum_{k\in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(t - y_k)\right). \tag{9}$$

However, the asymptotic behaviour of $\widehat{F}_{yc}(t)$ is linked to the selection of the vector $\mathbf{t_g}$, and therefore the optimum choice of this vector is needed. Under simple random sampling, previous analyses have considered the optimal selection for $\mathbf{t_g}$ [9,23,24,39]. Thus, [24] stated both the optimal dimension and the optimal selection of the vector $\mathbf{t_{opt}}(t)$ that minimises the asymptotic variance of $\widehat{F}_{yc}(t)$, and thus defined a new estimator based on the optimal vector, $\widehat{F}_{ycopt}(t)$.

In the next section, based on the approach described by [25], we propose estimators for the distribution function that incorporate both quantitative and qualitative variables, whilst seeking to avoid high dimensionality in the auxiliary information. In this proposal, the calibration weights are calculated using a projection of the auxiliary information onto a low-dimension Euclidean space, using the MDS procedure together with an appropriate dissimilarity measure to address the auxiliary obtained from categorical variables.

3. Calibration estimators of $\mathbf{F}_y(t)$ based on multidimensional scaling

3.1. Multidimensional scaling based on the auxiliary information $\mathbf{x_k}$

As mentioned above, survey statisticians today have ready access to large databases and it is quite common for these to include both quantitative and qualitative variables. Given an auxiliary vector $\mathbf{x}_k$, we assume that its dimension J is a large value and, moreover, that a set of $L < J$ variables included in the auxiliary vector $\mathbf{x}_k$ are qualitative variables. To obtain the pseudo-variable g from the approach presented by [19], we must consider the representation of the qualitative variables through the corresponding dummy variables. Thus, if we assume without loss of generality that the variables $x_1, x_2, \dots, x_L$ are qualitative and that $F_l \geq 2$ denotes the number of different categories or levels in the lth variable, with $l = 1, \dots, L$, then to avoid perfect multicollinearity,

we must consider $F_l - 1$ dummy variables in order to incorporate x_l into the definition of the pseudo-variable g. Therefore, given the variable x_l, we consider the dummy variables corresponding to its first $F_l - 1$ categories, that is, the dummy variables I_{j_l} with $j = 1, \dots, F_l - 1$ given by:

$$I_{j_l k} = \begin{cases} 1 & \text{if unit } k \text{ has the category } j_l \\ 0 & \text{otherwise} \end{cases}$$

Now, if we denote by A_{jk} the vector $(I_{1_l k}, I_{2_l k}, \dots, I_{(F_l-1)_l k})$, for $l = 1, \dots, L$, we can replace the original auxiliary vector $\mathbf{x}_k$ by a new auxiliary vector given by

$$\mathbf{z}_k = (A_{1k}, \dots, A_{Lk}, x_{(L+1)k}, \dots, x_{Pk}) \quad \text{for } k \in U$$

The dimension of the new vector $\mathbf{z}_k$ is given by $M = J + Q - 2 \cdot L$ where

$$Q = \sum_{l=1}^{L} F_l \geq 2 \cdot L$$

so the new dimension $M > J$. Hence, it is very likely that when using the new auxiliary vector $\mathbf{z}_k$ to obtain the pseudo-variable g, there will be multicollinearity among the auxiliary variables.

To avoid this multicollinearity, instead of considering the representation of the qualitative variables $x_1, x_2, \dots, x_L$ through their corresponding dummy vectors $A_{1k}, \dots, A_{Lk}$, we now discuss an alternative approach based on the proposal by [25].

For this purpose, consider a new set of auxiliary variables, obtained from the multidimensional scaling application with the original vector $\mathbf{x}_k$. This multidimensional scaling procedure is based on the similarity matrix calculated from that defined by Gower [40].

To introduce Gower's similarity measure, we assume that among the L qualitative variables, L_1 are binary and L_2 are non-binary, so that $L_1 + L_2 = L$. Consequently, $L_3 = J - L = J - L_1 - L_2$ quantitative variables are included in the auxiliary vector $\mathbf{x}_k$. For two units $i, k \in U$, the similarity index proposed by Gower is given by:

$$S_{ik} = \frac{b_1 + b_2 + \sum_{l=1}^{L_3} \left(1 - \frac{|x_{li} - x_{lk}|}{R_l}\right)}{(L_1 - \bar{b}_1) + L_2 + L_3} \tag{10}$$

where b_1 and $\bar{b}_1$ denote the positive and negative matches, respectively, for the L_1 binary variables, b_2 denotes the matches for the L_2 non-binary qualitative variables and R_l is the range of the lth quantitative variable.

From [40], the $N \times N$ similarity matrix $\mathbf{S}$ which entries S_{ik} for all pairs of units $i, k \in U$ is positive semi-definite. This is a relevant property because it allows us to represent the matrix $\mathbf{S}$ as a set of points in a multidimensional Euclidean space [41]. To do so, we consider the measure of distance given by:

$$D_{ik} = \sqrt{2(1 - S_{ik})}$$

and the following matrix

$$\mathbf{B} = -\frac{1}{2}\mathbf{EFE} = \mathbf{ESE}$$

with $\mathbf{F} = D_{ik}^2$ and

$$\mathbf{E} = \mathbf{I} - \frac{1}{N} \cdot \mathbf{1}\mathbf{1}'$$

where $rank(\mathbf{B}) = H \leq N - 1$.

Since the matrix $\mathbf{S}$ is positive semidefinite, the distance matrix $\mathbf{D}$ is Euclidean and B is positive semidefinite [41]. If we consider the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_H > 0$ of $\mathbf{B}$, the matrix

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}^{1/2}$$

is a $N \times H$ matrix whose associated Euclidean distance matrix is $\mathbf{D}$, where

$$\mathbf{V} = (v_{(1)}, \dots, v_{(h)})$$

is a $N \times H$ matrix with the eigenvectors $v_{(i)}$ associated with the positive eigenvalues of $\mathbf{B}$ by column such that $v'_{(i)} v_{(i)} = \lambda_i$ and $\mathbf{\Lambda}$ is a $H \times H$ diagonal matrix with the positive eigenvalues of $\mathbf{B}$ [41].

In the first alternative approach to developing a new calibration estimator for $F_y(t)$, if we denote by $\mathbf{c}_k$ for $k \in U$, the vector of dimension H from the matrix $\mathbf{C}$, the dimension of $\mathbf{c}_k$ can be reduced by multidimensional scaling. To do so, we take the $h \leq H$ largest eigenvalues of $\mathbf{B}$, and then define the following matrix:

$$\mathbf{C}_h = \mathbf{V}_h \mathbf{\Lambda}_h^{1/2}$$

where $\mathbf{V}_h$ is a $N \times h$ matrix that contains the eigenvectors $v_{(i)}$ associated with the h largest eigenvalues of $\mathbf{B}$ by column and $\mathbf{\Lambda}_h$ is a $h \times h$ diagonal matrix with largest eigenvalues of $\mathbf{B}$ [41].

Now, with the auxiliary vector $\mathbf{c}_k^h$ of dimension h for $k \in U$ extracted from $\mathbf{C}_h$ we can define the following pseudo-variable:

$$g_k^* = \widehat{\beta}_c' \mathbf{c}_k^h \text{ for } k = 1, 2, \dots, N$$

$$\widehat{\beta}_c = \left(\sum_{k \in s} d_k \mathbf{c}_k^h (\mathbf{c}_k^h)' \right)^{-1} \cdot \sum_{k \in s} d_k \mathbf{c}_k^h y_k. \tag{11}$$

The new set of auxiliary variables is related to the original set of qualitative or mixed variables through the Euclidean distances, but at the same time we maintain the implicit assumption of linearity [25] in constructing the pseudo-variable.

Thus, a new calibration estimator can be obtained for $F_y(t)$ by minimising (6) under the condition:

$$\frac{1}{N} \sum_{k \in s} \omega_k \Delta(\boldsymbol{\nu}_{\mathbf{g}^*} - g_k^*) = F_{g^*}(\boldsymbol{\nu}_{\mathbf{g}^*}) \tag{12}$$

where $F_{g^*}(\boldsymbol{\nu}_{\mathbf{g}^*})$ denotes the distribution function for g^* evaluated at $\boldsymbol{\nu}_{\mathbf{g}^*} = (\nu_1, \dots, \nu_P)'$ and where $\nu_j \quad j = 1, 2, \dots, P$ are points chosen such that $\nu_1 < \nu_2 < \cdots \nu_P$.

If we assume that the following matrix Γ is nonsingular:

$$\Gamma = \sum_{k \in s} d_k q_k \Delta(\boldsymbol{\nu}_{\mathbf{g}^*} - g_k^*) \Delta(\boldsymbol{\nu}_{\mathbf{g}^*} - g_k^*)'$$

the calibration estimator obtained is given by:

$$\widehat{F}_{ymds1}(t) = \widehat{F}_{YHT}(t) + \left(F_{g^*}(\boldsymbol{\nu}_{\mathbf{g}^*}) - \widehat{F}_{G^*HT}(\boldsymbol{\nu}_{\mathbf{g}^*}) \right)' \cdot \widehat{\Theta}(\boldsymbol{\nu}_{\mathbf{g}^*}) \tag{13}$$

where

$$\widehat{\Theta}(\boldsymbol{\nu}_{\mathbf{g}^*}) = \Gamma^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\boldsymbol{\nu}_{\mathbf{g}^*} - g_k^*) \Delta(t - y_k)$$

and $\widehat{F}_{G^*HT}(\boldsymbol{\nu}_{\mathbf{g}^*})$ is the Horvitz–Thompson estimator of F_{g^*} at $\boldsymbol{\nu}_{\mathbf{g}^*}$.

The resulting estimator $\widehat{F}_{ymds1}(t)$ is asymptotically unbiased and the asymptotic variance is [19]:

$$AV(\widehat{F}_{ymds1}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (d_k U_k)(d_l U_l) \tag{14}$$

where $U_k = \Delta(t - y_k) - \Delta(\boldsymbol{\nu}_{\mathbf{g}^*} - g_k^*)' \cdot \Theta(\boldsymbol{\nu}_{\mathbf{g}^*})$, with

$$\Theta(\boldsymbol{\nu}_{\mathbf{g}^*}) = \left(\sum_{k \in U} q_k \Delta(\boldsymbol{\nu}_{\mathbf{g}^*} - g_k^*) \Delta(\boldsymbol{\nu}_{\mathbf{g}^*} - g_k^*)' \right)^{-1} \cdot \left(\sum_{k \in U} q_k \Delta(\boldsymbol{\nu}_{\mathbf{g}^*} - g_k^*) \Delta(t - y_k) \right). \tag{15}$$

As the asymptotic behaviour of $\widehat{F}_{ymds1}(t)$ depends on the vector $\boldsymbol{\nu}_{\mathbf{g}^*}$, under simple random sampling, we can consider the optimal vector $\mathbf{t}_{\mathbf{opt}}(t)$ from [24] or its reduced version from [37].

3.2. Multidimensional scaling based on the complete information related to the auxiliary distribution functions

To take advantage of all available auxiliary information, we now consider a second alternative.

In this case, for all variables z_{mk} with $m = 1, \dots, M$ in the auxiliary vector $\mathbf{z}_k$, we can define the N-dimensional auxiliary vector given by:

$$(\boldsymbol{\tau}_k^m)' = \Big(\Delta(z_{m1} - z_{mk}), \dots, \Delta(z_{mN} - z_{mk}) \Big), \quad m = 1, \dots, M$$

Next, with the M auxiliary vectors $(\boldsymbol{\tau}_k^m)$, $m = 1, \dots, M$, we can define the following $N \cdot M$-dimensional vector:

$$\boldsymbol{\Upsilon}_k' = ((\boldsymbol{\tau}_k^1)', \dots, (\boldsymbol{\tau}_k^M)'). \tag{16}$$

Although all the information about the distribution functions of the variables incorporated in the auxiliary vector $\mathbf{z}_k$ are embraced in the auxiliary vector $\boldsymbol{\Upsilon}_k$, we cannot calibrate an estimator for $F_y(t)$ with $\boldsymbol{\Upsilon}_k$ since this would generate a large number of calibration conditions, many of which might be incompatible, or otherwise produce over-calibration.

Once again, multidimensional scaling can be used to reduce the dimension of $\boldsymbol{\Upsilon}_k$ with Gower's similarity measure, given by:

$$S_{ik} = \frac{\sum_{m=1}^{M} \sum_{l=1}^{N} \Big(1 - |\Delta(z_{ml} - z_{mi}) - \Delta(z_{ml} - z_{mk})| \Big)}{MN} \tag{17}$$

The derived distance $D_{ik} = \sqrt{2(1 - S_{ik})}$ is equivalent to the distance from Manhattan and we consider the matrix $\mathbf{S}_Y$ and $\mathbf{D}_Y$ with the entries S_{ik} and D_{ik}, respectively. As in the previous cases, the following matrix:

$$\mathbf{B}_Y = -\frac{1}{2} \mathbf{E} \mathbf{D}_Y \mathbf{E} = \mathbf{E} \mathbf{S}_Y \mathbf{E}$$

is positive semi-definite. If we consider that $rank(\mathbf{B}_Y) = J \leq N-1$ with the j largest egienvalues, we can obtain the matrix $\mathbf{C}_j$ as in the previous cases and the corresponding auxiliary vector $\mathbf{r}_k = \mathbf{c}_k^j$ for all population unit $k \in U$.

By minimising (6) under the constraints:

$$\frac{1}{N}\sum_{k\in s}\omega_k \mathbf{r}_k = \frac{1}{N}\sum_{k\in U}\mathbf{r}_k = \bar{R} \tag{18}$$

we can obtain a new calibration estimator $\widehat{F}_{ymds2}(t)$ for $F_y(t)$ given by

$$\widehat{F}_{ymds2}(t) = \widehat{F}_{YHT}(t) + \left(\bar{R} - \bar{R}_{HT}\right)' \cdot \widehat{Q} \tag{19}$$

where we assume that the matrix $\mathbf{\Psi}$:

$$\mathbf{\Psi} = \sum_{k\in s} d_k q_k \mathbf{r}_k \cdot \mathbf{r}_k'$$

is non-singular and

$$\widehat{Q} = \mathbf{\Psi}^{-1} \cdot \sum_{k\in s} d_k q_k \mathbf{r}_k \Delta(t - y_k)$$

and $\bar{R}_{HT}$ is the Horvitz–Thompson estimator for $\bar{R}$.

Following [19], the estimator $\widehat{F}_{ymds2}(t)$ is asymptotically unbiased and its asymptotic variance is:

$$AV(\widehat{F}_{ymds2}(t)) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k U_k)(d_l U_l) \tag{20}$$

where $U_k = \Delta(t - y_k) - \mathbf{r}_k' \cdot Q$, with

$$Q = \left(\sum_{k\in U} q_k \mathbf{r}_k \cdot \mathbf{r}_k'\right)^{-1} \cdot \left(\sum_{k\in U} q_k \mathbf{r}_k \Delta(t - y_k)\right). \tag{21}$$

3.3. Multidimensional scaling based on the auxiliary vector $\mathbf{x_k}$ and the auxiliary distribution functions

Let us now consider an alternative, incorporating the auxiliary information from the auxiliary vector $\mathbf{x_k}$ and the auxiliary distribution functions associated with the ordered qualitative and quantitative variables included in $\mathbf{x_k}$. Previously, to do so, given the auxiliary vector $\mathbf{x}_k$, we assumed that the variables $x_1, x_2, \ldots, x_L$ were qualitative and that $x_{L+1}, x_2, \ldots, x_J$ were quantitative. Among the qualitative attributes, F_l denoted the number of different categories or levels in the lth variable, with $l = 1, \ldots, L$ and we assumed that L_1 were binary variables and L_2 were non-binary variables. Now, for the L_1 binary variables, we assume that L_{1N} are qualitative variables with non-ordered categories and L_{1O} are qualitative variables with ordered attributes with $L_{1N} + L_{1O} = L_1$. Similarly, for the L_2 non-binary variables, we assume that L_{2N} are qualitative variables with non-ordered categories and L_{2O} are qualitative variables with ordered attributes with $L_{2N} + L_{2O} = L_2$.

For the L_{1O} binary qualitative variables with ordered attributes, for each unit $k \in U$, we consider the vector $F1O_k$ with the distribution function values associated:

$$F1O_k = (F1O_{1k}, \ldots, F1O_{L_{1O}k})' \tag{22}$$

where

$$F1O_{lk} = \frac{1}{N}\sum_{i\in U}\Delta(x_{lk} - x_{li}), \quad l = 1, \ldots, L_{1O}.$$

In a similar way, for the L_{2O} non-binary qualitative variables with ordered attributes, for each unit $k \in U$, we consider the vector $F2O_k$ given by:

$$F2O_k = (F2O_{1k}, \ldots, F2O_{L_{2O}k})' \tag{23}$$

where

$$F2O_{lk} = \frac{1}{N}\sum_{i\in U}\Delta(x_{lk} - x_{li}), \quad l = 1, \ldots, L_{2O}.$$

Finally, for the L_3 quantitative variables x_l we also consider for each $k \in U$ the vector with the distribution function values:

$$F_k = (F_{1k}, \ldots, F_{L_3 k})' \tag{24}$$

with

$$F_{lk} = \frac{1}{N}\sum_{i\in U}\Delta(x_{lk} - x_{li}), \quad l = 1, \ldots, L_3.$$

Now, for every unit $k \in U$, we can consider the following auxiliary vector

$$\mathbf{T}_k' = (\mathbf{x}_k', F1O_k', F2O_k', F_k)$$

It is clear that the dimension of $\mathbf{T}'_k$ is

$$L_1 + L_2 + L_3 + L_{1O} + L_{2O} + L_3 = 2J - L_{1N} - L_{2N}.$$

The dimension of the new vector $\mathbf{T}'_k$ can be reduced with multidimensional scaling. To do so, we obtain the similarity matrix $\mathbf{S_T}$ with entries S_{ij} calculated with Gower's similarity index:

$$S_{ik} = \frac{b_1 + b_2 + \sum_{l=1}^{L_3}\left(1 - \frac{|x_{li} - x_{lk}|}{R_l}\right) + \sum_{l=1}^{L_{1O}}\left(1 - \frac{|F1O_{li} - F1O_{lk}|}{R_l}\right)}{(L_1 - \bar{b}_1) + L_{1O} + L_2 + L_{2O} + 2L_3} + \frac{\sum_{l=1}^{L_{2O}}\left(1 - \frac{|F2O_{li} - F2O_{lk}|}{R_l}\right) + \sum_{l=1}^{L_3}\left(1 - \frac{|F_{li} - F_{lk}|}{R_l}\right)}{(L_1 - \bar{b}_1) + L_{1O} + L_2 + L_{2O} + 2L_3} \tag{25}$$

and the corresponding distance matrix $\mathbf{D_T}$.

As in the previous cases, with the u largest eigenvalues from the matrix

$$\mathbf{B_T} = -\frac{1}{2}\mathbf{E}\mathbf{F_T}\mathbf{E} = \mathbf{E}\mathbf{S_T}\mathbf{E}$$

we can obtain the matrix $\mathbf{C}_u$ in the usual way that for every unit $k \in U$ contains an auxiliary vector $\mathbf{m}_k = \mathbf{c}_k^u$.

A new calibration estimator $\widehat{F}_{ymds3}(t)$ can be obtained by minimising (6) under the following conditions:

$$\frac{1}{N}\sum_{k\in s}\omega_k\mathbf{m}_k = \frac{1}{N}\sum_{k\in U}\mathbf{m}_k = \bar{M}. \tag{26}$$

If we denote the Horvitz–Thompson estimator for $\bar{M}$ by $\bar{M}_{HT}$, the expression for $\widehat{F}_{ymds3}(t)$ is given as follows:

$$\widehat{F}_{ymds3}(t) = \widehat{F}_{YHT}(t) + \left(\bar{M} - \bar{M}_{HT}\right)' \cdot \widehat{W} \tag{27}$$

with

$$\widehat{W} = \chi^{-1} \cdot \sum_{k\in s} d_k q_k \mathbf{m}_k \Delta(t - y_k)$$

where we assume that the matrix χ defined as:

$$\chi = \sum_{k\in s} d_k q_k \mathbf{m}_k \cdot \mathbf{m}'_k$$

is non-singular.

Using the linearity properties of the calibration estimator given in [15], it can be obtained that the estimator $\widehat{F}_{ymds3}(t)$ is asymptotically unbiased with an asymptotic variance given by the following expression:

$$AV(\widehat{F}_{ymds3}(t)) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k\varepsilon_k)(d_l\varepsilon_l) \tag{28}$$

where $\varepsilon_k = \Delta(t - y_k) - \mathbf{m}'_k \cdot K$, with

$$K = \left(\sum_{k\in U} q_k \mathbf{m}_k \cdot \mathbf{m}'_k\right)^{-1} \cdot \left(\sum_{k\in U} q_k \mathbf{m}_k \Delta(t - y_k)\right). \tag{29}$$

4. Properties of the calibration estimators based on multidimensional scaling

When a new estimator $\widehat{F}_y(t)$ of the distribution function $F_y(t)$ is introduced, it is important to determine whether $\widehat{F}_y(t)$ is also a distribution function, that is, whether $\widehat{F}_y(t)$ satisfies the following properties:

(i) $\widehat{F}_y(t)$ is continuous on the right,
(ii) (a) $\lim_{t\to-\infty}\widehat{F}_y(t) = 0$ and (b) $\lim_{t\to+\infty}\widehat{F}_y(t) = 1$,
(iii) $\widehat{F}_y(t)$ is monotone nondecreasing.

Compliance with the above properties allows us to estimate population quantiles and wage inequality measures based on quantiles, through the inverse function of the estimator $\widehat{F}_y(t)$ [8,42]. However, not all the new calibration estimators proposed satisfy all the properties of the distribution function.

The estimator $\widehat{F}_{ymds1}(t)$ always satisfies properties (i) and (iia) whereas properties (iib) and (iii) are respectively satisfied if a sufficiently large value of ν_P is selected in the vector $\boldsymbol{\nu}_{\mathbf{g}^*}$ and if we choose $q_k = c$ for all $k \in U$.

Both $\widehat{F}_{ymds2}(t)$ and $\widehat{F}_{ymds3}(t)$ satisfy conditions (i) and (iia). Additionally, both of them satisfy condition (iib) if the following constraint

$$\frac{1}{N}\sum_{k\in s}\omega_k = 1 \tag{30}$$

is added to the respective calibration processes (18) and (26). Henceforth, we assume the inclusion of condition (30) in the calibration processes to obtain the respective estimators $\widehat{F}_{ymds2}(t)$ and $\widehat{F}_{ymds3}(t)$.

Finally, the estimators $\widehat{F}_{ymds2}(t)$ and $\widehat{F}_{ymds3}(t)$ satisfy property (iii) if and only if the respective calibrated weights ω_k are positive for all $k \in s$. With the chi-square distance (6), we cannot guarantee positive calibrated weights for all sample units, but the distance function associated with the raking method avoids negative calibrated weights [43]. Specifically, the distance based on the raking method is given as follows:

$$G_s(\omega_k, d_k) = \sum_{k\in s}\frac{1}{q_k}\left(\omega_k \log\frac{\omega_k}{d_k} - \omega_k + d_k\right). \tag{31}$$

The raking distance is especially recommended when we wish to calibrate with respect to qualitative auxiliary information or if we wish to calibrate for known cell counts or known marginal counts in a frequency table of any dimension [25,43] and this can be a useful option for satisfying property (iii) with the estimators $\widehat{F}_{ymds2}(t)$ and $\widehat{F}_{ymds3}(t)$.

5. Simulation study

In this section, we discuss a simulation study conducted to compare the performance of the proposed estimators $\widehat{F}_{ymds1}(t)$; $\widehat{F}_{ymds2}(t)$ and $\widehat{F}_{ymds3}(t)$. To analyse these estimators, the simulation study was carried out applying specific procedures developed in R [version 4.3.1]. In addition, alternative estimators of the distribution function $F_y(t)$ were included. These alternative estimators were the Horvitz–Thompson estimator $\widehat{F}_{HT}$ and the following indirect estimators, the Chambers–Dunstan estimator [44] $\widehat{F}_{CD}(t)$, the Kovar–Mantel Rao-Estimator [45] $\widehat{F}_{RKM}(t)$ and the calibration estimator $\widehat{F}_{yc}(t)$ proposed by [19], for which two alternatives were considered. Denoting by $Q_g(\alpha)$ the quantile of variable g of order α, we considered the calibration estimators: $\widehat{F}^1_{yc}(t)$, with auxiliary vector $\mathbf{t_g} = (Q_g(0.5))$ and $\widehat{F}^3_{yc}(t)$ with auxiliary vector $\mathbf{t_g} = (Q_g(0.25), Q_g(0.5), Q_g(0.75))$. All of these indirect estimators employ the pseudo-variable g based on the auxiliary vector $\mathbf{z}$ that includes the representation of the qualitative information through dummy variables. For the proposed estimator $\widehat{F}_{ymds1}(t)$, we also included two versions, $\widehat{F}^1_{ymds1}(t)$ based on $\boldsymbol{\nu}_{\mathbf{g}^*} = (Q_{g^*}(0.5))$ and $\widehat{F}^3_{ymds1}(t)$ based on $\boldsymbol{\nu}_{\mathbf{g}^*} = (Q_{g^*}(0.25), Q_{g^*}(0.5), Q_{g^*}(0.75))$.

The simulation study encompassed three populations, one of which is real and the rest, simulated.

The first population is a generated population of size $N = 500$ called SPANISH500. The population includes the variables age, nationality, gender, weight and access to the Internet. These variables were generated such that the final population was similar to the Spanish population pyramid. The study variable is defined as follows:

$$y_k = 3 + 5\cdot \text{Internet} + \text{Age}/5 + \epsilon_k$$

where the values ϵ_k are independent identically distributed random variables with $\epsilon_k \sim N(0, 0.1)$.

The second population is a simulated population called SIMPOPULATION. The population size is $N = 1000$ and it includes 16 variables based on the procedure described in [46]. The first variable η_k was generated using independent and identically distributed values from a uniform distribution in $(0, 1)$. The other variables were generated from the following regression models:

$$\begin{aligned}
m_{1k} &= 1 + 2(\eta_k - 0.5) + \varepsilon_{1k};\ \varepsilon_{1k} \sim N(0, 0.01)\\
m_{2k} &= 1 + 2(\eta_k - 0.5) + \varepsilon_{2k};\ \varepsilon_{2k} \sim N(0, 0.04)\\
n_{1k} &= 1 + 2(\eta_k - 0.5)^2 + \zeta_{1k};\ \zeta_{1k} \sim N(0, 0.01)\\
n_{2k} &= 1 + 2(\eta_k - 0.5)^2 + \zeta_{2k};\ \zeta_{2k} \sim N(0, 0.04)\\
n_{3k} &= 1 + 2(\eta_k - 0.5)^2 + \zeta_{3k};\ \zeta_{2k} \sim N(0, 0.1)\\
n_{4k} &= 1 + 2(\eta_k - 0.5)^2 + \zeta_{4k};\ \zeta_{2k} \sim N(0, 0.04)\\
b_{1k} &= 1 + 2(\eta_k - 0.5) + \exp\left(-200(\eta_k - 0.5)^2\right) + \gamma_{1k};\ \gamma_{1k} \sim N(0, 0.01)\\
b_{2k} &= 1 + 2(\eta_k - 0.5) + \exp\left(-200(\eta_k - 0.5)^2\right) + \gamma_{2k};\ \gamma_{2k} \sim N(0, 0.04)\\
b_{3k} &= 1 + 2(\eta_k - 0.5) + \exp\left(-200(\eta_k - 0.5)^2\right) + \gamma_{3k};\ \gamma_{3k} \sim N(0, 0.1)\\
b_{4k} &= 1 + 2(\eta_k - 0.5) + \exp\left(-200(\eta_k - 0.5)^2\right) + \gamma_{4k};\ \gamma_{4k} \sim N(0, 0.4)\\
e_{1k} &= \exp(-8\eta_k) + \tau_{1k};\ \tau_{1k} \sim N(0, 0.01)\\
e_{2k} &= \exp(-8\eta_k) + \tau_{2k};\ \tau_{2k} \sim N(0, 0.04)\\
e_{3k} &= \exp(-8\eta_k) + \tau_{3k};\ \tau_{3k} \sim N(0, 0.1)\\
e_{4k} &= \exp(-8\eta_k) + \tau_{4k};\ \tau_{4k} \sim N(0, 0.4)\\
c_{1k} &= 2 + \sin(2\pi\eta_k) + \rho_{1k};\ \rho_{1k} \sim N(0, 0.01)\\
c_{2k} &= 2 + \sin(2\pi\eta_k) + \rho_{2k};\ \rho_{2k} \sim N(0, 0.01)
\end{aligned}$$

Table 1
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SPANISH500, simple random sampling and $PE = 50\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0024	1	0.0042	1	0.0017	1	0.0013	1
$\widehat{F}_{CD}$	0.0084	0.9263	0.0063	0.9456	0.0045	0.9393	0.0034	0.9462
$\widehat{F}_{RKM}$	0.0037	1.0130	0.0035	1.0094	0.0016	1.0016	0.0015	1.0003
$\widehat{F}^{1}_{yc}$	0.0011	1.5237	0.0042	1.5522	0.0013	1.5480	0.0012	1.5873
$\widehat{F}^{3}_{yc}$	0.0020	1.4429	0.0060	1.4164	0.0018	1.3934	0.0012	1.4382
$\widehat{F}^{1}_{ymds1}$	0.0051	1.1109	0.0036	1.0724	0.0015	1.1096	0.0045	1.0971
$\widehat{F}^{3}_{ymds1}$	0.0044	0.9082	0.0044	0.8231	0.0022	0.8658	0.0029	0.7937
$\widehat{F}_{ymds2}$	0.0107	0.4714	0.0078	0.4681	0.0013	0.5008	0.0022	0.2944
$\widehat{F}_{ymds3}$	0.0037	0.7565	0.0085	0.7323	0.0029	0.7536	0.0012	0.7364

The study variable is c_{2k} and the remaining variables are considered in the auxiliary vector $\mathbf{x}_k$. The variables m_{2k}, n_{2k}, n_{4k} and b_{2k} are divided into two categories via the median, and the variables m_{1k}, n_{1k}, n_{3k}, b_{1k} and b_{3k} via the quartiles, into four categories.

The last population considered was the dataset EUSILC from the R package "laeken". This population is synthetically generated from the European Union Statistics on Income and Living Conditions in Austria. The dataset has 14 827 observations and 27 variables. The study variable was employee cash or near cash income (var $py010n$) and the remaining variables were included in the auxiliary vector $\mathbf{x}_k$.

The selection criterion for the dimension of principal coordinates obtained through multidimensional scaling is based on the goodness of fit measure (GOF) [41] given by:

$$GOF(h) = \frac{\sum_{k=1}^{h} \lambda_k}{\sum_{k=1}^{H} |\lambda_k|} \cdot 100 \tag{32}$$

In all populations, the following percentages $PE = 50\%$, 60%, 70% and 80% were considered. Thus, in each case the minimum number of principal coordinates h was retained so that the $GOF(h)$ value was greater than or equal to the PE value considered. For each percentage value PE, 1000 different samples were drawn by simple random sampling without replacement for four different sizes. With each sample, estimates of the distribution function $F(t)$ at 11 points were obtained with all the estimators included in the simulation study. The 11 points considered were the quantiles $Q_y(\alpha)$ for $\alpha = 0.1$, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8 and 0.9.

The measures employed to compare the performance of each estimator included in the simulation study were the average relative bias (AVRB) and the average relative efficiency (AVRE), respectively given by:

$$\text{AVRB}(t) = \frac{1}{11}\sum_{q=1}^{11} |\text{RB}(t_q)|, \quad \text{AVRE}(t) = \frac{1}{11}\sum_{q=1}^{11} \text{RE}(t_q)$$

with

$$\text{RB}(t) = \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{F}(t)_b - F_y(t)}{F_y(t)} \quad \text{and} \quad \text{RE}(t) = \frac{MSE[\widehat{F}(t)]}{MSE[\widehat{F}_{HT}(t)]}, \tag{33}$$

and $MSE[\widehat{F}(t)]$ denotes the empirical mean square error for $\widehat{F}(t)$ defined as follows:

$$MSE[\widehat{F}(t)] = B^{-1}\sum_{b=1}^{B} [\widehat{F}(t)_b - F_y(t)]^2$$

where b indexes the bth simulation run, $\widehat{F}(t)$ is an estimator for the distribution function and $MSE[\widehat{F}_{HT}(t)]$ is the empirical mean square error for the Horvitz–Thompson estimator.

This simulation study is implemented in the statistical computing environment R using code developed by the authors. This code is available from the authors on request.

The results for the first simulation study with the SPANISH500 population are summarised in Tables 1, 2, 3 and 4 for the different values of PE.

As can be seen, all the estimators perform well in terms of relative bias, and none is uniformly better than any other. Concerning efficiency, in all cases $\widehat{F}_{ymds2}$ outperforms the other estimators, especially with regard to $\widehat{F}_{HT}$. Moreover, the estimators $\widehat{F}_{ymds3}$ and $\widehat{F}^{3}_{ymds1}$ are more efficient than $\widehat{F}_{HT}$. The remaining indirect estimators are less efficient than $\widehat{F}_{HT}$, with the exception of $\widehat{F}_{CD}$, which is slightly better. Another aspect of interest is that for $PE = 80$ (except when size $n = 125$), the efficiency of estimator $\widehat{F}^{3}_{ymds1}$ is

Table 2
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SPANISH500, simple random sampling and $PE = 60\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{Y}}_{HT}$	0.0146	1	0.0049	1	0.0030	1	0.0037	1
$\widehat{\bar{Y}}_{CD}$	0.0166	0.9115	0.0077	0.9381	0.0031	0.9585	0.0047	0.9630
$\widehat{\bar{Y}}_{RKM}$	0.0173	1.0124	0.0056	1.0132	0.0027	1.0095	0.0043	1.0146
$\widehat{\bar{Y}}^{1}_{yc}$	0.0170	1.6334	0.0052	1.5832	0.0030	1.5235	0.0036	1.5857
$\widehat{\bar{Y}}^{3}_{yc}$	0.0164	1.4899	0.0047	1.4428	0.0027	1.4223	0.0034	1.5075
$\widehat{\bar{Y}}^{1}_{ymds1}$	0.0175	1.1115	0.0085	1.1030	0.0046	1.1189	0.0072	1.0915
$\widehat{\bar{Y}}^{3}_{ymds1}$	0.0154	0.9149	0.0051	0.8387	0.0029	0.8067	0.0055	0.8302
$\widehat{\bar{Y}}_{ymds2}$	0.0051	0.4874	0.0027	0.4758	0.0069	0.4752	0.0018	0.2902
$\widehat{\bar{Y}}_{ymds3}$	0.0079	0.7665	0.0038	0.7445	0.0037	0.7355	0.0028	0.7398

Table 3
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SPANISH500, simple random sampling and $PE = 70\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{Y}}_{HT}$	0.0050	1	0.0032	1	0.0017	1	0.0013	1
$\widehat{\bar{Y}}_{CD}$	0.0069	0.9227	0.0065	0.9393	0.0045	0.9415	0.0044	0.9592
$\widehat{\bar{Y}}_{RKM}$	0.0052	1.0168	0.0031	1.0099	0.0014	1.0043	0.0021	1.0064
$\widehat{\bar{Y}}^{1}_{yc}$	0.0051	1.4904	0.0036	1.4625	0.0015	1.5336	0.0023	1.5872
$\widehat{\bar{Y}}^{3}_{yc}$	0.0050	1.4257	0.0052	1.3873	0.0016	1.4251	0.0021	1.4318
$\widehat{\bar{Y}}^{1}_{ymds1}$	0.0075	1.0654	0.0073	1.0800	0.0038	1.0980	0.0019	1.0736
$\widehat{\bar{Y}}^{3}_{ymds1}$	0.0066	0.8538	0.0057	0.8705	0.0023	0.8315	0.0021	0.8090
$\widehat{\bar{Y}}_{ymds2}$	0.0079	0.3294	0.0017	0.3141	0.0010	0.3029	0.0016	0.3113
$\widehat{\bar{Y}}_{ymds3}$	0.0065	0.7758	0.0033	0.7622	0.0018	0.7553	0.0030	0.7260

Table 4
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SPANISH500, simple random sampling and $PE = 80\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{Y}}_{HT}$	0.0038	1	0.0094	1	0.0025	1	0.0037	1
$\widehat{\bar{Y}}_{CD}$	0.0063	0.9244	0.0097	0.9501	0.0048	0.9449	0.0035	0.9553
$\widehat{\bar{Y}}_{RKM}$	0.0024	1.0185	0.0092	1.0099	0.0031	1.0079	0.0013	1.0044
$\widehat{\bar{Y}}^{1}_{yc}$	0.0038	1.5763	0.0091	1.5534	0.0026	1.5553	0.0013	1.5282
$\widehat{\bar{Y}}^{3}_{yc}$	0.0057	1.4531	0.0112	1.3963	0.0030	1.4258	0.0015	1.4536
$\widehat{\bar{Y}}^{1}_{ymds1}$	0.0035	0.8653	0.0059	0.8816	0.0030	0.8417	0.0048	1.1002
$\widehat{\bar{Y}}^{3}_{ymds1}$	0.0127	0.3830	0.0059	0.3540	0.0067	0.3481	0.0021	0.8139
$\widehat{\bar{Y}}_{ymds2}$	0.0086	0.2830	0.0070	0.2601	0.0031	0.2537	0.0012	0.2965
$\widehat{\bar{Y}}_{ymds3}$	0.0052	0.8853	0.0054	0.7063	0.0027	0.7016	0.0019	0.7513

considerably higher, while the estimator $\widehat{\bar{Y}}^{1}_{ymds1}$ is more efficient than $\widehat{\bar{Y}}_{CD}$ and $\widehat{\bar{Y}}_{HT}$. Finally, the proposed estimators $\widehat{\bar{Y}}^{1}_{ymds1}$, $\widehat{\bar{Y}}^{3}_{ymds1}$, generally achieve better values for AVRE than their respective versions based on the usual calibration.

For the second simulation study with the SIMPOPULATION, Tables 5, 6, 7 and 8 summarise the results for the four values of PE.

As with the previous populations, the results for the SIMPOPULATION show that there is no estimator that minimises the bias in a uniform way. However, a notable bias reduction is achieved by $\widehat{\bar{Y}}_{ymds3}$ for the cases $PE = 60$ and $PE = 80$ with size $n = 125$.

Concerning efficiency, the estimators $\widehat{\bar{Y}}_{ymds2}$, $\widehat{\bar{Y}}_{ymds3}$ and $\widehat{\bar{Y}}^{3}_{ymds1}$ present a notable improvement over $\widehat{\bar{Y}}_{HT}$. In general, the estimators $\widehat{\bar{Y}}_{ymds2}$ and $\widehat{\bar{Y}}_{ymds3}$ uniformly present the lowest values of AVRE. The estimator $\widehat{\bar{Y}}^{1}_{ymds1}$ is also more efficient than $\widehat{\bar{Y}}_{HT}$ in all cases. Among the other indirect estimators, only $\widehat{\bar{Y}}_{CD}$ is more efficient than $\widehat{\bar{Y}}_{HT}$, and then only slightly so. Moreover, this estimator always performs worse than $\widehat{\bar{Y}}_{ymds2}$, $\widehat{\bar{Y}}_{ymds3}$, $\widehat{\bar{Y}}^{3}_{ymds1}$ and $\widehat{\bar{Y}}^{1}_{ymds1}$. As in the previous populations, the proposed estimators $\widehat{\bar{Y}}^{1}_{ymds1}$, $\widehat{\bar{Y}}^{3}_{ymds1}$ present lower values of AVRE than their respective usual calibration versions $\widehat{\bar{Y}}^{1}_{yc}$, $\widehat{\bar{Y}}^{3}_{yc}$.

Finally, the results for the EUSILC population, for all values of PE, are summarised in Tables 9, 10, 11 and 12.

Table 5
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SIMPOPULATION, simple random sampling and $PE = 50\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0015	1	0.0036	1	0.0021	1	0.0031	1
$\widehat{F}_{CD}$	0.0092	0.9365	0.0091	0.9457	0.0068	0.9548	0.0057	0.9489
$\widehat{F}_{RKM}$	0.0010	1.0033	0.0042	1.0144	0.0022	1.0088	0.0032	1.0040
$\widehat{F}^{1}_{yc}$	0.0016	1.5062	0.0015	1.5478	0.0021	1.5053	0.0033	1.5155
$\widehat{F}^{3}_{yc}$	0.0015	1.3917	0.0026	1.4138	0.0027	1.3488	0.0029	1.3624
$\widehat{F}^{1}_{ymds1}$	0.0020	0.8989	0.0012	0.9495	0.0022	0.9272	0.0032	0.8718
$\widehat{F}^{3}_{ymds1}$	0.0035	0.4462	0.0022	0.4774	0.0025	0.4381	0.0033	0.4171
$\widehat{F}_{ymds2}$	0.0041	0.4943	0.0044	0.4965	0.0026	0.4785	0.0017	0.2938
$\widehat{F}_{ymds3}$	0.0020	0.3396	0.0015	0.3531	0.0019	0.3302	0.0022	0.3084

Table 6
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SIMPOPULATION, simple random sampling and $PE = 60\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0044	1	0.0013	1	0.0043	1	0.0026	1
$\widehat{F}_{CD}$	0.0112	0.9369	0.0088	0.9384	0.0067	0.94082	0.0091	0.9523
$\widehat{F}_{RKM}$	0.0042	1.0112	0.0010	1.0008	0.0048	1.0029	0.0026	1.0025
$\widehat{F}^{1}_{yc}$	0.0049	1.4890	0.0010	1.5656	0.0042	1.4757	0.0026	1.5667
$\widehat{F}^{3}_{yc}$	0.0050	1.4024	0.0013	1.4039	0.0047	1.3751	0.0025	1.4179
$\widehat{F}^{1}_{ymds1}$	0.0038	0.8879	0.0036	0.9000	0.0025	0.8927	0.0026	0.8758
$\widehat{F}^{3}_{ymds1}$	0.0041	0.3740	0.0022	0.4023	0.0038	0.3661	0.0027	0.4297
$\widehat{F}_{ymds2}$	0.0055	0.3153	0.0029	0.3236	0.0008	0.3085	0.0028	0.3079
$\widehat{F}_{ymds3}$	0.0049	0.3202	0.0024	0.3263	0.0006	0.3120	0.0028	0.3146

Table 7
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SIMPOPULATION, simple random sampling and $PE = 70\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0042	1	0.0033	1	0.0030	1	0.0021	1
$\widehat{F}_{CD}$	0.0094	0.9456	0.0084	0.9441	0.0099	0.9430	0.0082	0.9559
$\widehat{F}_{RKM}$	0.0038	1.0115	0.0033	1.0109	0.0029	1.0028	0.0020	1.0021
$\widehat{F}^{1}_{yc}$	0.0040	1.5081	0.0036	1.5142	0.0029	1.5529	0.0019	1.5258
$\widehat{F}^{3}_{yc}$	0.0040	1.3974	0.0040	1.3832	0.0029	1.3859	0.0022	1.3653
$\widehat{F}^{1}_{ymds1}$	0.0056	0.8793	0.0033	0.8744	0.0029	0.8881	0.0022	0.8899
$\widehat{F}^{3}_{ymds1}$	0.0033	0.3898	0.0047	0.3772	0.0029	0.3850	0.0015	0.4339
$\widehat{F}_{ymds2}$	0.0057	0.3075	0.0013	0.3020	0.0031	0.3053	0.0028	0.3070
$\widehat{F}_{ymds3}$	0.0051	0.2889	0.0024	0.2804	0.0037	0.2775	0.0024	0.3140

The estimator $\widehat{F}_{CD}$ presents obvious problems of bias and efficiency in all cases. All other estimators present good results for bias. For efficiency, the best estimators are clearly $\widehat{F}_{ymds2}$ and $\widehat{F}_{ymds3}$, followed by $\widehat{F}^{3}_{ymds1}$ and $\widehat{F}^{1}_{ymds1}$. Only these four estimators are more efficient than $\widehat{F}_{HT}$, except in the cases of $PE = 80$ and $n = 75$, where $\widehat{F}^{1}_{ymds1}$ performs worse than $\widehat{F}_{HT}$.

From the results derived from the three simulation studies carried out, we conclude that in general the efficiency of estimator $\widehat{F}_{ymds1}$ is considerably improved when the calibration process is considered with three points (estimator $\widehat{F}^{3}_{ymds1}$) rather than a single point (estimator $\widehat{F}^{1}_{ymds1}$). Additionally, it is not necessary to consider a high value of PE to achieve a notable improvement in efficiency, since with $PE = 50\%$ a considerable reduction in the $AVRE$ is achieved with the proposed estimators. In fact, in general, the improvement in efficiency with $PE = 50\%$ remains stable with higher values of PE, with the sole exception of the estimator $\widehat{F}_{ymds2}$, in which case the efficiency is substantially higher for $PE = 80$ in all populations except EUSILC. This efficiency improvement for $PE = 80$ is also shown by the estimators $\widehat{F}^{1}_{ymds1}$ and $\widehat{F}^{3}_{ymds1}$ but only for the SPANISH500 population.

Table 8
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SIMPOPULATION, simple random sampling and $PE = 80\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0015	1	0.0031	1	0.0043	1	0.0039	1
$\widehat{F}_{CD}$	0.0092	0.9366	0.0097	0.9452	0.0067	0.9408	0.0082	0.9560
$\widehat{F}_{RKM}$	0.0010	1.0033	0.0026	1.0051	0.0048	1.0029	0.0043	1.0100
$\widehat{F}^{1}_{yc}$	0.0016	1.5056	0.0029	1.5273	0.0042	1.4757	0.0029	1.5239
$\widehat{F}^{3}_{yc}$	0.0014	1.3918	0.0028	1.3691	0.0047	1.3751	0.0034	1.4123
$\widehat{F}^{1}_{ymds1}$	0.0040	0.8761	0.0029	0.8733	0.0030	0.8991	0.0015	0.9345
$\widehat{F}^{3}_{ymds1}$	0.0060	0.3996	0.0044	0.3958	0.0052	0.3756	0.0023	0.4422
$\widehat{F}_{ymds2}$	0.0050	0.2603	0.0032	0.2384	0.0013	0.2386	0.0022	0.3150
$\widehat{F}_{ymds3}$	0.0038	0.2954	0.0030	0.2776	0.0008	0.2712	0.0018	0.3213

Table 9
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population EUSILC, simple random sampling and $PE = 50\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0023	1	0.0005	1	0.0011	1	0.0013	1
$\widehat{F}_{CD}$	0.1821	9.8648	0.1810	13.2486	0.1799	17.0118	0.1832	21.1746
$\widehat{F}_{RKM}$	0.0024	1.0141	0.0006	1.0373	0.0016	1.0198	0.0012	0.9916
$\widehat{F}^{1}_{yc}$	0.0040	1.4453	0.0008	1.5029	0.0010	1.4942	0.0009	1.4107
$\widehat{F}^{3}_{yc}$	0.0036	1.2670	0.0005	1.3270	0.0012	1.3331	0.0010	1.2541
$\widehat{F}^{1}_{ymds1}$	0.0037	0.8738	0.0008	0.9674	0.0011	0.8759	0.0015	0.9486
$\widehat{F}^{3}_{ymds1}$	0.0017	0.6855	0.0010	0.7308	0.0019	0.6752	0.0008	0.6900
$\widehat{F}_{ymds2}$	0.0006	0.3265	0.0007	0.3251	0.0019	0.3199	0.0009	0.3225
$\widehat{F}_{ymds3}$	0.0015	0.3942	0.0006	0.3946	0.0015	0.3889	0.0006	0.3840

Table 10
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population EUSILC, simple random sampling and $PE = 60\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0010	1	0.0014	1	0.0009	1	0.0005	1
$\widehat{F}_{CD}$	0.1841	9.7379	0.1807	12.8360	0.1797	16.5563	0.1835	19.3017
$\widehat{F}_{RKM}$	0.0008	1.0272	0.0017	1.0226	0.0014	1.0232	0.0007	1.0093
$\widehat{F}^{1}_{yc}$	0.0027	1.5920	0.0018	1.4387	0.0012	1.4551	0.0005	1.4627
$\widehat{F}^{3}_{yc}$	0.0014	1.3313	0.0013	1.2632	0.0015	1.2744	0.0012	1.2942
$\widehat{F}^{1}_{ymds1}$	0.0019	0.9521	0.0019	0.9488	0.0010	0.8931	0.0017	0.8892
$\widehat{F}^{3}_{ymds1}$	0.0008	0.7043	0.0010	0.7362	0.0009	0.6780	0.0011	0.6521
$\widehat{F}_{ymds2}$	0.0009	0.3286	0.0003	0.3195	0.0005	0.3286	0.0009	0.3139
$\widehat{F}_{ymds3}$	0.0012	0.3857	0.0005	0.3943	0.0006	0.3869	0.0007	0.3727

5.1. Simulation studies with Midzuno sampling

To illustrate the robustness of the proposed estimators against the sampling design used, the simulations were repeated in all populations but obtaining the samples through Midzuno sampling.

For the SPANISH500 population, Midzuno sampling was performed considering the variable *Age*. Tables 13, 14, 15 and 16 summarise the results obtained for the same sample size n and PE values considered previously.

Tables 13, 14, 15 and 16 show that the relative bias results obtained with Midzuno sampling differ from those of simple random sampling in that the best-performing estimators in most cases are $\widehat{F}_{CD}$ and $\widehat{F}_{RKM}$, although $\widehat{F}_{ymds2}$ and $\widehat{F}_{ymds3}$ present very similar values. Indeed, the latter has the lowest relative bias of all for $PE = 70$ and $n = 75, 100, 125$. In general, $\widehat{F}^{3}_{ymds1}$ presents higher levels of bias than $\widehat{F}_{ymds2}$ and $\widehat{F}_{ymds3}$, but as the PE value increases, this bias decreases sharply, such that for $PE = 80$, it is less than that recorded for $\widehat{F}_{ymds2}$ and $\widehat{F}_{ymds3}$. The estimator $\widehat{F}^{1}_{ymds1}$ performs better than $\widehat{F}_{HT}$ and the usual calibrated estimators $\widehat{F}^{1}_{yc}$ and $\widehat{F}^{3}_{yc}$.

Concerning relative efficiency, all the indirect estimators perform better than $\widehat{F}_{HT}$, and $\widehat{F}_{ymds2}$ is by far the best in all cases. The estimator $\widehat{F}_{ymds3}$ is also more efficient than the other estimators except in the case of $PE = 80$, where $\widehat{F}_{CD}$, $\widehat{F}_{RKM}$ and $\widehat{F}^{3}_{ymds1}$ present

Table 11
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population EUSILC, simple random sampling and $PE = 70\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0027	1	0.0014	1	0.0011	1	0.0011	1
$\widehat{F}_{CD}$	0.1832	9.9963	0.1824	13.0813	0.1800	15.9325	0.1839	19.8261
$\widehat{F}_{RKM}$	0.0025	1.0004	0.0019	1.0255	0.0013	1.0158	0.0015	0.9947
$\widehat{F}^{1}_{yc}$	0.0032	1.4711	0.0004	1.5680	0.0011	1.4756	0.0009	1.4387
$\widehat{F}^{3}_{yc}$	0.0031	1.3031	0.0011	1.3098	0.0022	1.2555	0.0010	1.2312
$\widehat{F}^{1}_{ymds1}$	0.0013	0.9541	0.0007	0.9542	0.0005	0.9070	0.0013	0.9094
$\widehat{F}^{3}_{ymds1}$	0.0009	0.7258	0.0022	0.7308	0.0021	0.6819	0.0008	0.6696
$\widehat{F}_{ymds2}$	0.0006	0.3440	0.0006	0.3395	0.0004	0.3194	0.0007	0.3102
$\widehat{F}_{ymds3}$	0.0012	0.4155	0.0006	0.4057	0.0004	0.3906	0.0005	0.3828

Table 12
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population EUSILC, simple random sampling and $PE = 80\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.0009	1	0.0013	1	0.0010	1	0.0008	1
$\widehat{F}_{CD}$	0.1833	10.3742	0.1800	13.7810	0.1796	16.9243	0.1834	20.4737
$\widehat{F}_{RKM}$	0.0010	1.0207	0.0004	1.0300	0.0012	1.0108	0.0008	1.0116
$\widehat{F}^{1}_{yc}$	0.0008	1.5386	0.0023	1.5078	0.0010	1.5612	0.0007	1.4983
$\widehat{F}^{3}_{yc}$	0.0008	1.2853	0.0019	1.2928	0.0009	1.3077	0.0007	1.2999
$\widehat{F}^{1}_{ymds1}$	0.0025	0.9613	0.0029	1.0067	0.0023	0.9684	0.0020	0.8411
$\widehat{F}^{3}_{ymds1}$	0.0012	0.7384	0.0015	0.7267	0.0029	0.6982	0.0024	0.5950
$\widehat{F}_{ymds2}$	0.0007	0.3346	0.0008	0.3455	0.0011	0.3476	0.0007	0.1887
$\widehat{F}_{ymds3}$	0.0009	0.4070	0.0013	0.4224	0.0018	0.4148	0.0008	0.2594

Table 13
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SPANISH500, Midzuno sampling and $PE = 50\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9890	1.0000	0.9753	1.0000	0.9561	1.0000	0.9312	1.0000
$\widehat{F}_{CD}$	0.0089	0.0374	0.0124	0.0251	0.0132	0.0187	0.0150	0.0151
$\widehat{F}_{RKM}$	0.0116	0.0414	0.0139	0.0270	0.0135	0.0199	0.0147	0.0160
$\widehat{F}^{1}_{yc}$	0.4735	0.2700	0.4734	0.2630	0.4618	0.2544	0.4449	0.2461
$\widehat{F}^{3}_{yc}$	0.2481	0.1054	0.2445	0.0905	0.2411	0.0832	0.2312	0.0771
$\widehat{F}^{1}_{ymds1}$	0.2508	0.1015	0.2548	0.0944	0.2509	0.0903	0.2472	0.0894
$\widehat{F}^{3}_{ymds1}$	0.1129	0.0437	0.1243	0.0357	0.1295	0.0326	0.1346	0.0315
$\widehat{F}_{ymds2}$	0.0155	0.0261	0.0167	0.0161	0.0158	0.0127	0.0158	0.0096
$\widehat{F}_{ymds3}$	0.0097	0.0375	0.0156	0.0236	0.0174	0.0177	0.0184	0.0139

better efficiency. As with the results for bias, the efficiency of $\widehat{F}^{3}_{ymds1}$ improves as the value of PE increases, and for $PE = 80$ it is only outperformed by $\widehat{F}_{ymds2}$. The estimator $\widehat{F}^{1}_{ymds1}$ always achieves much higher efficiency than $\widehat{F}_{HT}$ and $\widehat{F}^{1}_{yc}$ and in general it also surpasses the efficiency of $\widehat{F}^{3}_{yc}$.

Midzuno sampling was then applied to the SIMPOPULATION, using the variable b_{4k}. Tables 17, 18, 19 and 20 show the results for the usual sample sizes n and for the usual PE values.

These Tables 17, 18, 19 and 20 also show that no estimator uniformly presents a lower bias, although in most cases $\widehat{F}_{ymds2}$ and $\widehat{F}_{ymds3}$ present the lowest relative bias, specially for high values of PE, results that are only surpassed, in some circumstances, by $\widehat{F}_{RKM}$. As in the previous cases, the bias obtained by estimator $\widehat{F}^{3}_{ymds1}$ is considerably less than that recorded for $\widehat{F}^{1}_{yc}$, $\widehat{F}^{3}_{yc}$ and $\widehat{F}_{HT}$. Moreover, $\widehat{F}^{1}_{ymds1}$ also presents lower values for bias than $\widehat{F}^{1}_{yc}$, $\widehat{F}^{3}_{yc}$ and $\widehat{F}_{HT}$

All estimators considerably improve the efficiency of $\widehat{F}_{HT}$, but $\widehat{F}_{ymds2}$, $\widehat{F}_{ymds3}$ and $\widehat{F}^{3}_{ymds1}$ are outstanding in this respect, with $\widehat{F}_{ymds2}$ uniformly presenting the best performance. The estimator $\widehat{F}^{1}_{ymds1}$ achieves better efficiency than the usual calibrated estimators $\widehat{F}^{1}_{yc}$ and $\widehat{F}^{3}_{yc}$.

Table 14
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SPANISH500, Midzuno sampling and $PE = 60\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9892	1.0000	0.9755	1.0000	0.9561	1.0000	0.9311	1.0000
$\widehat{F}_{CD}$	0.0153	0.0386	0.0134	0.0249	0.0148	0.0190	0.0111	0.0149
$\widehat{F}_{RKM}$	0.0173	0.0424	0.0147	0.0270	0.0145	0.0200	0.0113	0.0156
$\widehat{F}^{1}_{yc}$	0.4880	0.2794	0.4798	0.2675	0.4627	0.2555	0.4400	0.2429
$\widehat{F}^{3}_{yc}$	0.2651	0.1111	0.2553	0.0944	0.2401	0.0827	0.2255	0.0754
$\widehat{F}^{1}_{ymds1}$	0.2511	0.1007	0.2461	0.0914	0.2391	0.0855	0.2321	0.0832
$\widehat{F}^{3}_{ymds1}$	0.0799	0.0432	0.0753	0.0298	0.0727	0.0255	0.0671	0.0222
$\widehat{F}_{ymds2}$	0.0250	0.0280	0.0175	0.0170	0.0166	0.0120	0.0101	0.0097
$\widehat{F}_{ymds3}$	0.0231	0.0380	0.0199	0.0233	0.0165	0.0169	0.0117	0.0134

Table 15
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SPANISH500, Midzuno sampling and $PE = 70\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9891	1.0000	0.9753	1.0000	0.9559	1.0000	0.9309	1.0000
$\widehat{F}_{CD}$	0.0089	0.0396	0.0125	0.0254	0.0105	0.0181	0.0120	0.0149
$\widehat{F}_{RKM}$	0.0115	0.0431	0.0136	0.0272	0.0125	0.0190	0.0135	0.0159
$\widehat{F}^{1}_{yc}$	0.4920	0.2855	0.4738	0.2613	0.4561	0.2501	0.4393	0.2428
$\widehat{F}^{3}_{yc}$	0.2582	0.1114	0.2508	0.0912	0.2353	0.0802	0.2248	0.0750
$\widehat{F}^{1}_{ymds1}$	0.2437	0.0991	0.2420	0.0911	0.2368	0.0856	0.2340	0.0842
$\widehat{F}^{3}_{ymds1}$	0.0577	0.0430	0.0553	0.0291	0.0521	0.0224	0.0510	0.0191
$\widehat{F}_{ymds2}$	0.0311	0.0224	0.0275	0.0139	0.0326	0.0104	0.0324	0.0086
$\widehat{F}_{ymds3}$	0.0127	0.0423	0.0096	0.0243	0.0055	0.0171	0.0044	0.0146

Table 16
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SPANISH500, Midzuno sampling and $PE = 80\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9892	1.0000	0.9755	1.0000	0.9561	1.0000	0.9311	1.0000
$\widehat{F}_{CD}$	0.0153	0.0386	0.0134	0.0249	0.0148	0.0190	0.0111	0.0149
$\widehat{F}_{RKM}$	0.0173	0.0424	0.0147	0.0270	0.0145	0.0200	0.0113	0.0156
$\widehat{F}^{1}_{yc}$	0.4880	0.2794	0.4798	0.2675	0.4627	0.2555	0.4400	0.2429
$\widehat{F}^{3}_{yc}$	0.2651	0.1111	0.2553	0.0944	0.2401	0.0827	0.2255	0.0754
$\widehat{F}^{1}_{ymds1}$	0.1800	0.0782	0.1731	0.0692	0.1703	0.0643	0.1640	0.0617
$\widehat{F}^{3}_{ymds1}$	0.0366	0.0209	0.0336	0.0133	0.0334	0.0103	0.0303	0.0086
$\widehat{F}_{ymds2}$	0.0429	0.0212	0.0358	0.0118	0.0336	0.0088	0.0319	0.0072
$\widehat{F}_{ymds3}$	0.0380	0.0623	0.0448	0.0319	0.0487	0.0220	0.0576	0.0203

Finally, for the EUSILC population, Midzuno sampling was performed considering the variable Age and results are displayed in Tables 21, 22, 23 and 24.

From the results shown in Tables 21, 22, 23 and 24, we conclude that the estimator that most reduces the bias is $\widehat{F}_{ymds3}$ in all cases, followed by $\widehat{F}_{ymds2}$. The estimator $\widehat{F}^{3}_{ymds1}$ also shows less bias than the other estimators except for $PE = 70, 80$ where it is slightly surpassed by $\widehat{F}_{RKM}$. Regarding relative efficiency, as in the previous case, all the indirect estimators are more efficient than $\widehat{F}_{HT}$. Estimators $\widehat{F}_{ymds2}$ and $\widehat{F}_{ymds3}$ are notably more efficient than the other estimators. Likewise, $\widehat{F}^{3}_{ymds1}$ is generally more efficient than the other estimators, while $\widehat{F}_{CD}$ obtains the worst results of all the indirect estimators.

5.2. Variability of the set of calibration weights

In this subsection, we further analyse the performance of the proposed estimators by focusing on the variability of the final set of calibration weights for each of the estimators considered and also for the usual calibration estimator. For each of the three populations analysed using the Midzuno sampling design, we now measure the variability of the calibrated weights in each of

Table 17
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SIMPOPULATION, Midzuno sampling and $PE = 50\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9976	1.0000	0.9945	1.0000	0.9901	1.0000	0.9845	1.0000
$\widehat{F}_{CD}$	0.0096	0.0360	0.0135	0.0238	0.0105	0.0181	0.0091	0.0140
$\widehat{F}_{RKM}$	0.0052	0.0398	0.0051	0.0259	0.0035	0.0193	0.0035	0.0149
$\widehat{F}^{1}_{yc}$	0.4738	0.2516	0.4808	0.2510	0.4742	0.2429	0.4707	0.2401
$\widehat{F}^{3}_{yc}$	0.2168	0.0821	0.2269	0.0738	0.2205	0.0656	0.2191	0.0618
$\widehat{F}^{1}_{ymds1}$	0.1568	0.0788	0.1604	0.0693	0.1598	0.0644	0.1591	0.0617
$\widehat{F}^{3}_{ymds1}$	0.0295	0.0221	0.0318	0.0150	0.0352	0.0115	0.0314	0.0095
$\widehat{F}_{ymds2}$	0.0026	0.0170	0.0056	0.0101	0.0037	0.0073	0.0033	0.0056
$\widehat{F}_{ymds3}$	0.0024	0.0177	0.0062	0.0105	0.0054	0.0075	0.0047	0.0057

Table 18
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SIMPOPULATION, Midzuno sampling and $PE = 60\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9976	1.0000	0.9944	1.0000	0.9901	1.0000	0.9845	1.0000
$\widehat{F}_{CD}$	0.0117	0.0370	0.0111	0.0228	0.0093	0.0173	0.0081	0.0147
$\widehat{F}_{RKM}$	0.0035	0.0409	0.0067	0.0250	0.0011	0.0184	0.0010	0.0154
$\widehat{F}^{1}_{yc}$	0.4772	0.2537	0.4721	0.2427	0.4744	0.2426	0.4724	0.2416
$\widehat{F}^{3}_{yc}$	0.2195	0.0827	0.2181	0.0689	0.2221	0.0664	0.2194	0.0625
$\widehat{F}^{1}_{ymds1}$	0.1612	0.0785	0.1587	0.0680	0.1585	0.0644	0.1565	0.0617
$\widehat{F}^{3}_{ymds1}$	0.0325	0.0218	0.0304	0.0145	0.0303	0.0119	0.0287	0.0100
$\widehat{F}_{ymds2}$	0.0068	0.0167	0.0028	0.0100	0.0015	0.0077	0.0015	0.0060
$\widehat{F}_{ymds3}$	0.0056	0.0173	0.0028	0.0104	0.0025	0.0080	0.0023	0.0061

Table 19
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SIMPOPULATION, Midzuno sampling and $PE = 70\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9976	1.0000	0.9944	1.0000	0.9902	1.0000	0.9845	1.0000
$\widehat{F}_{CD}$	0.0087	0.0363	0.0076	0.0244	0.0112	0.0172	0.0100	0.0139
$\widehat{F}_{RKM}$	0.0053	0.0405	0.0070	0.0265	0.0054	0.0183	0.0026	0.0147
$\widehat{F}^{1}_{yc}$	0.4728	0.2515	0.4741	0.2450	0.4787	0.2476	0.4738	0.2434
$\widehat{F}^{3}_{yc}$	0.2190	0.0828	0.2174	0.0704	0.2267	0.0676	0.2233	0.0640
$\widehat{F}^{1}_{ymds1}$	0.1590	0.0782	0.1556	0.0686	0.1588	0.0639	0.1570	0.0611
$\widehat{F}^{3}_{ymds1}$	0.0302	0.0215	0.0298	0.0140	0.0299	0.0112	0.0300	0.0091
$\widehat{F}_{ymds2}$	0.0022	0.0168	0.0016	0.0097	0.0031	0.0072	0.0017	0.0053
$\widehat{F}_{ymds3}$	0.0017	0.0172	0.0019	0.0099	0.0033	0.0075	0.0026	0.0056

the calibrated estimators and that of each of the selected samples. Thus, for each PE value, for each sample size n and for each calibration estimator, we have 1000 measurements of the variability of the final set of weights. Figs. 1, 2, 3 and 4 show the boxplots of these 1000 measurements for each calibration estimator according to the PE value, organised by sample sizes for the SPANISH500 population. Figs. 5, 6, 7 and 8 and Figs. 9, 10, 11 and 12, respectively, show similar boxplots for the SIMPOPULATION and EUSILC populations.

For the SPANISH500 population, Figs. 1, 2, 3 and 4 show that, in general, there is a similar degree of variability of the weight system for each calibrated estimator, although $\widehat{F}_{ymds2}$ and $\widehat{F}_{ymds3}$ present less variability and the usual calibrated estimators $\widehat{F}^{1}_{yc}$ and $\widehat{F}^{3}_{yc}$ present greater variability. Moreover, as the sample size n increases, the variability of the set of weights of all the estimators decreases considerably, to the extent that the differences among estimators become almost imperceptible.

Regarding the SIMPOPULATION, Figs. 5, 6, 7 and 8 reflect a general situation that is very similar to that shown for the SPANISH500 population, but on this occasion the estimator $\widehat{F}^{1}_{ymds1}$ presents slightly greater variability in the weights than $\widehat{F}^{3}_{yc}$.

Table 20
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population SIMPOPULATION, Midzuno sampling and $PE = 80\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9976	1.0000	0.9945	1.0000	0.9901	1.0000	0.9845	1.0000
$\widehat{F}_{CD}$	0.0082	0.0370	0.0095	0.0236	0.0099	0.0184	0.0094	0.0130
$\widehat{F}_{RKM}$	0.0053	0.0405	0.0021	0.0254	0.0035	0.0199	0.0038	0.0139
$\widehat{F}^{1}_{yc}$	0.4762	0.2549	0.4762	0.2466	0.4735	0.2419	0.4685	0.2375
$\widehat{F}^{3}_{yc}$	0.2205	0.0843	0.2216	0.0709	0.2202	0.0656	0.2165	0.0600
$\widehat{F}^{1}_{ymds1}$	0.1572	0.0779	0.1581	0.0683	0.1579	0.0646	0.1581	0.0608
$\widehat{F}^{3}_{ymds1}$	0.0303	0.0209	0.0297	0.0144	0.0317	0.0117	0.0300	0.0096
$\widehat{F}_{ymds2}$	0.0011	0.0163	0.0018	0.0099	0.0035	0.0075	0.0018	0.0055
$\widehat{F}_{ymds3}$	0.0024	0.0165	0.0017	0.0102	0.0047	0.0078	0.0037	0.0057

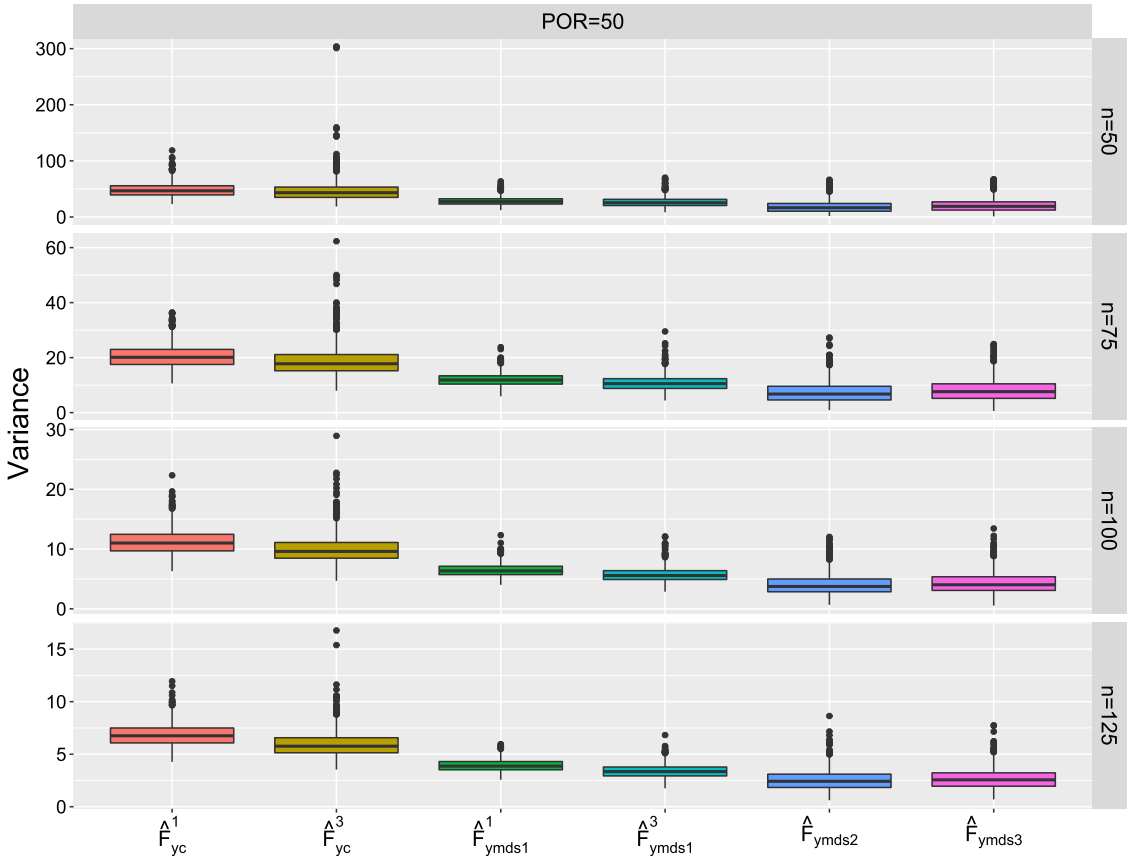

Fig. 1. Boxplot for variability of calibration weights SPANISH500 population, $PE = 50$.

Finally, for the EUSILC population, Figs. 9, 10, 11 and 12 show that the estimator $\widehat{F}_{ymds2}$ achieves the lowest variability, while $\widehat{F}^{3}_{ymds1}$ and $\widehat{F}_{ymds3}$ show the greatest variability in the set of weights. However, while the variability of $\widehat{F}^{3}_{ymds1}$ is less than that of the other estimators as the value of PE increases, the variability of $\widehat{F}_{ymds3}$ increases.

5.3. Simulation studies in subpopulations

To better understand the proposed estimators, in this subsection we analyse their performance with respect to the distribution function in different subpopulations. Specifically, we estimate the distribution function in the subpopulation of women from the SPANISH500 population (Tables 25, 26, 27 and 28) and the subpopulation of people with Internet access from the same population

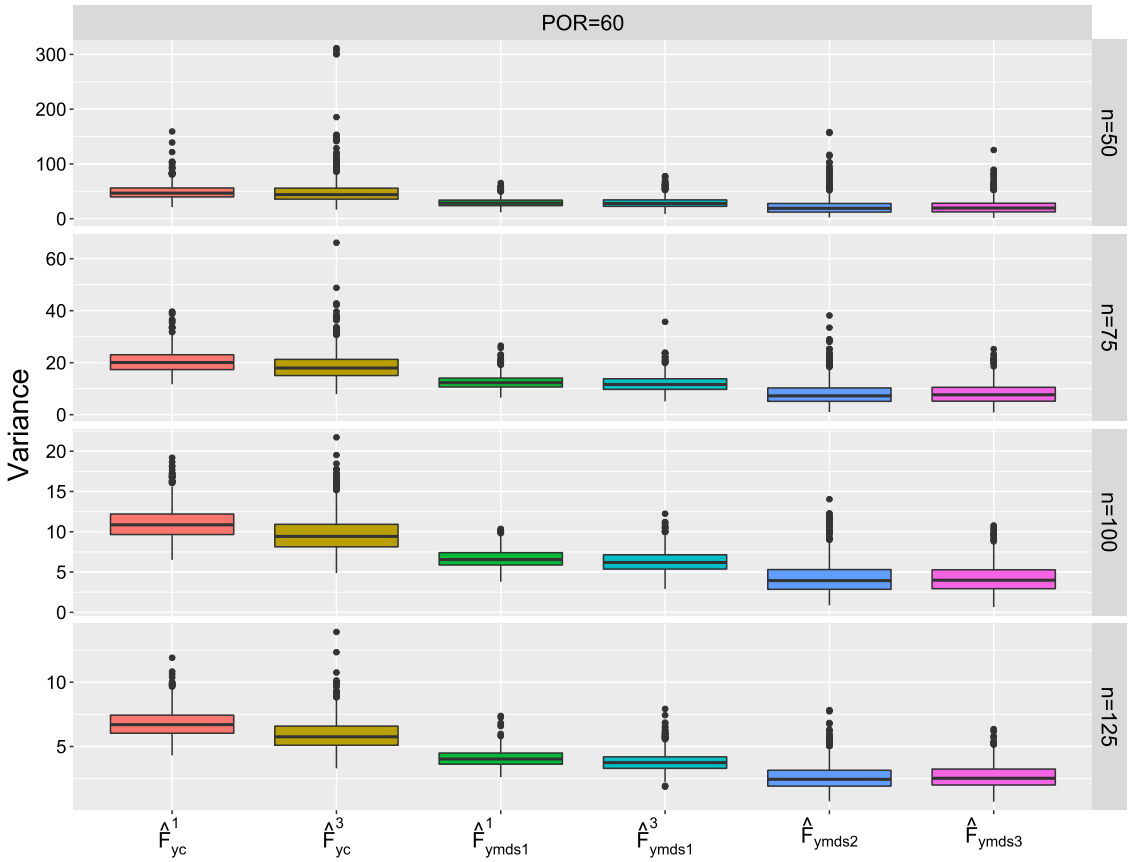

Fig. 2. Boxplot for variability of calibration weights SPANISH500 population, $PE = 60$.

Table 21
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population EUSILC, Midzuno sampling and $PE = 50\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\hat{F}_{HT}$	0.9988	1.0000	0.9978	1.0000	0.9965	1.0000	0.9950	1.0000
$\hat{F}_{CD}$	0.1883	0.0696	0.1883	0.0675	0.1856	0.0653	0.1856	0.0644
$\hat{F}_{RKM}$	0.0732	0.0120	0.0752	0.0108	0.0742	0.0099	0.0727	0.0091
$\hat{F}^1_{yc}$	0.1387	0.0357	0.1358	0.0345	0.1373	0.0347	0.1385	0.0346
$\hat{F}^3_{yc}$	0.0907	0.0130	0.0912	0.0120	0.0911	0.0115	0.0908	0.0110
$\hat{F}^1_{ymds1}$	0.1863	0.0579	0.1847	0.0574	0.1851	0.0576	0.1847	0.0575
$\hat{F}^3_{ymds1}$	0.0632	0.0114	0.0618	0.0106	0.0621	0.0102	0.0625	0.0098
$\hat{F}_{ymds2}$	0.0437	0.0036	0.0446	0.0034	0.0455	0.0032	0.0448	0.0030
$\hat{F}_{ymds3}$	0.0363	0.0038	0.0350	0.0031	0.0351	0.0028	0.0352	0.0025

(Tables 29, 30, 31 and 32). As in the previous simulation studies, we considered the same sample sizes and the same values for PE, and the sampling design considered was simple random sampling.

Regarding the results obtained in the subpopulation of women, no estimator was uniformly better than the rest in terms of relative bias. Although in most cases $\hat{F}^1_{yc}$, $\hat{F}^3_{yc}$ and $\hat{F}_{RKM}$ present the least bias, in certain situations ($n = 75{,}125$ for $PE = 70$), $\hat{F}_{ymds2}$ is the least biased. Regarding efficiency, again $\hat{F}_{ymds2}$ and $\hat{F}_{ymds3}$ perform best, especially the first of these, because for high values of PE, the results obtained by $\hat{F}_{ymds3}$ are slightly worse, and the efficiency is less than that obtained by $\hat{F}_{HT}$ for $PE = 80$ and $n = 50$, perhaps because this estimator begins to suffer from overcalibration. The estimators $\hat{F}^1_{ymds1}$ and $\hat{F}^3_{ymds1}$ are more efficient than $\hat{F}^1_{yc}$ and $\hat{F}^3_{yc}$, especially for large PE values, and even for $PE = 80$ their efficiency is greater than that of $\hat{F}_{CD}$ and $\hat{F}_{RKM}$.

Finally, for the subpopulation of people with Internet access, Tables 29, 30, 31 and 32 show that in general all estimators obtain similar values for relative bias, with the exception of $\hat{F}_{CD}$, which presents generalised problems in this respect, $\hat{F}_{ymds2}$ for low values

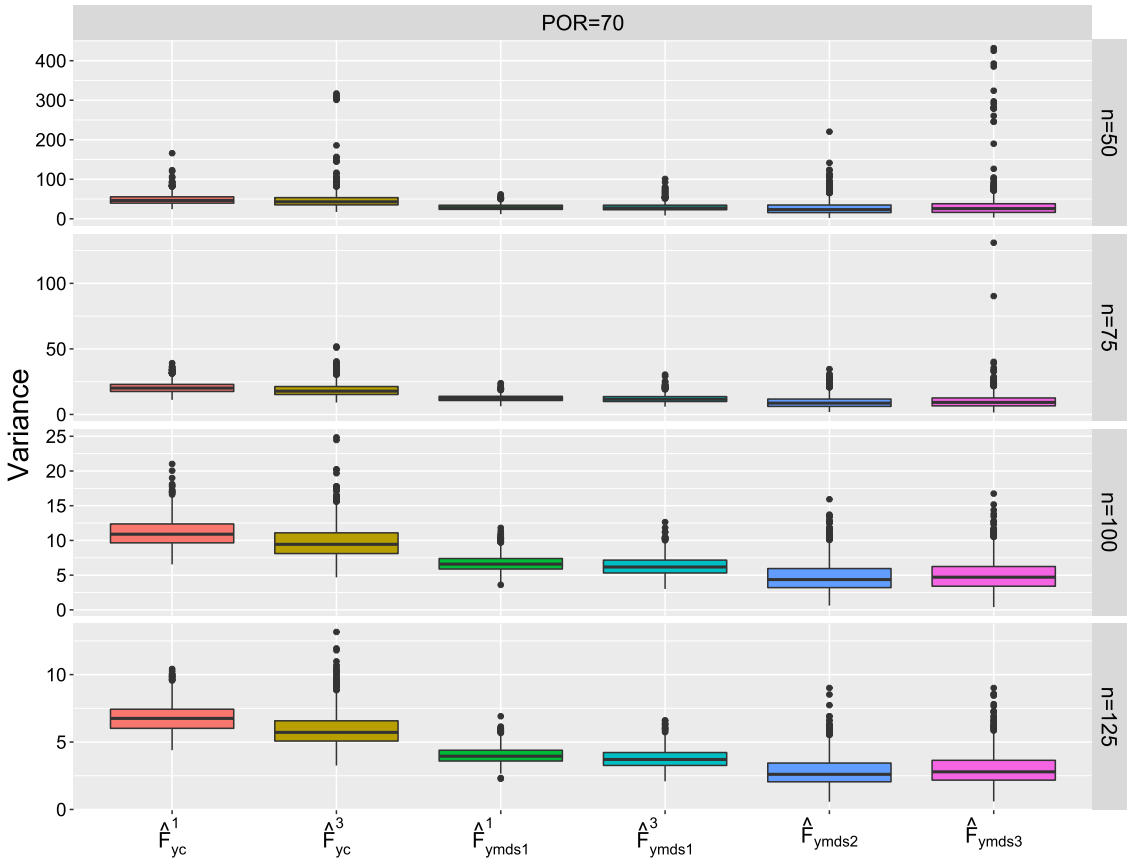

Fig. 3. Boxplot for variability of calibration weights SPANISH500 population, $PE = 70$.

Table 22
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population EUSILC, Midzuno sampling and $PE = 60\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\hat{F}_{HT}$	0.9988	1.0000	0.9978	1.0000	0.9965	1.0000	0.9950	1.0000
$\hat{F}_{CD}$	0.1887	0.0689	0.1890	0.0672	0.1867	0.0664	0.1852	0.0642
$\hat{F}_{RKM}$	0.0753	0.0120	0.0756	0.0107	0.0713	0.0093	0.0728	0.0090
$\hat{F}^1_{yc}$	0.1372	0.0356	0.1369	0.0346	0.1380	0.0347	0.1382	0.0345
$\hat{F}^3_{yc}$	0.0899	0.0127	0.0908	0.0120	0.0902	0.0113	0.0897	0.0107
$\hat{F}^1_{ymds1}$	0.1789	0.0562	0.1793	0.0563	0.1792	0.0562	0.1801	0.0564
$\hat{F}^3_{ymds1}$	0.0637	0.0104	0.0616	0.0094	0.0604	0.0090	0.0595	0.0085
$\hat{F}_{ymds2}$	0.0426	0.0037	0.0429	0.0034	0.0435	0.0032	0.0433	0.0030
$\hat{F}_{ymds3}$	0.0308	0.0041	0.0304	0.0031	0.0298	0.0027	0.0306	0.0024

of PE and small sample sizes and $\hat{F}_{ymds3}$ for $PE = 80$. No estimator uniformly achieves the lowest value for relative bias. As with the Women subpopulation, the estimator $\hat{F}_{ymds2}$ presents the greatest efficiency, followed by $\hat{F}^3_{ymds1}$ and $\hat{F}_{ymds3}$, but for $PE = 80$, the estimator $\hat{F}_{ymds3}$ shows signs of overcalibration, with efficiency values even lower than those obtained by $\hat{F}_{HT}$. And again as in the Women subpopulation, the estimator $\hat{F}^1_{ymds1}$ outperforms $\hat{F}^1_{yc}$ and $\hat{F}^3_{yc}$ and for $PE = 80$ it exceeds the efficiency of $\hat{F}_{CD}$ and $\hat{F}_{RKM}$.

6. Discussion

[25] first proposed the MDS technique for selecting appropriate variables for calibration weighting in surveys. This procedure can reduce the variance of the population estimator of the total or the mean of a variable, when survey calibration is used with auxiliary information that includes qualitative variables.

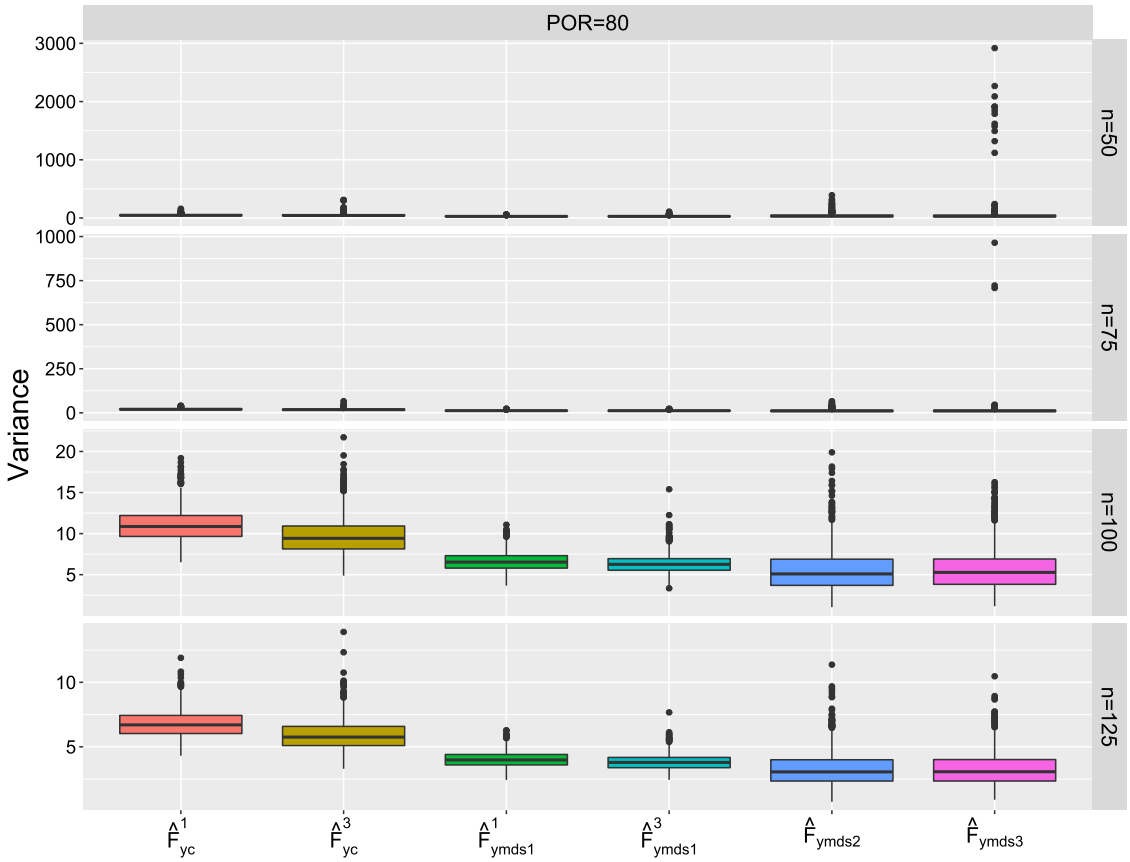

Fig. 4. Boxplot for variability of calibration weights SPANISH500 population, $PE = 80$.

Table 23
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population EUSILC, Midzuno sampling and $PE = 70\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9988	1.0000	0.9978	1.0000	0.9965	1.0000	0.9950	1.0000
$\widehat{F}_{CD}$	0.1883	0.0684	0.1891	0.0672	0.1862	0.0655	0.1853	0.0646
$\widehat{F}_{RKM}$	0.0766	0.0126	0.0751	0.0107	0.0742	0.0099	0.0720	0.0089
$\widehat{F}^1_{yc}$	0.1383	0.0356	0.1381	0.0350	0.1367	0.0346	0.1374	0.0345
$\widehat{F}^3_{yc}$	0.0908	0.0130	0.0899	0.0118	0.0900	0.0113	0.0902	0.0108
$\widehat{F}^1_{ymds1}$	0.1626	0.0545	0.1641	0.0546	0.1646	0.0548	0.1643	0.0549
$\widehat{F}^3_{ymds1}$	0.0801	0.0099	0.0774	0.0091	0.0789	0.0090	0.0777	0.0087
$\widehat{F}_{ymds2}$	0.0280	0.0026	0.0275	0.0023	0.0290	0.0021	0.0291	0.0020
$\widehat{F}_{ymds3}$	0.0266	0.0029	0.0243	0.0022	0.0263	0.0020	0.0255	0.0017

Based on this idea, we propose three alternatives to incorporate multidimensional scaling-based calibration [47] into the estimation of the distribution function. This approach provides a reliable alternative when mixed auxiliary information (i.e. with both qualitative and quantitative variables) is used to estimate the distribution function, compared to the usual procedure of incorporating qualitative variables through corresponding dummy variables. The methods we describe reduce the dimension of the auxiliary information and avoid overcalibration problems. Moreover, and unlike other alternatives based on principal components [38] that only admit quantitative auxiliary information, our proposal also facilitates the incorporation of qualitative auxiliary information. All of these proposals can be applied with any probability sampling design and provide a single set of calibrated weights that do not depend on the values of the study variable.

Table 24
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population EUSILC, Midzuno sampling and $PE = 80\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{F}_{HT}$	0.9988	1.0000	0.9978	1.0000	0.9965	1.0000	0.9950	1.0000
$\widehat{F}_{CD}$	0.1902	0.0704	0.1878	0.0677	0.1874	0.0653	0.1851	0.0639
$\widehat{F}_{RKM}$	0.0716	0.0115	0.0735	0.0108	0.0753	0.0101	0.0741	0.0095
$\widehat{F}^{1}_{yc}$	0.1403	0.0360	0.1385	0.0351	0.1374	0.0344	0.1371	0.0345
$\widehat{F}^{3}_{yc}$	0.0881	0.0127	0.0897	0.0119	0.0904	0.0113	0.0905	0.0110
$\widehat{F}^{1}_{ymds1}$	0.1596	0.0531	0.1597	0.0531	0.1597	0.0531	0.1593	0.0532
$\widehat{F}^{3}_{ymds1}$	0.0770	0.0094	0.0777	0.0090	0.0762	0.0086	0.0762	0.0084
$\widehat{F}_{ymds2}$	0.0128	0.0012	0.0136	0.0010	0.0129	0.0008	0.0125	0.0007
$\widehat{F}_{ymds3}$	0.0077	0.0018	0.0094	0.0014	0.0087	0.0011	0.0089	0.0009

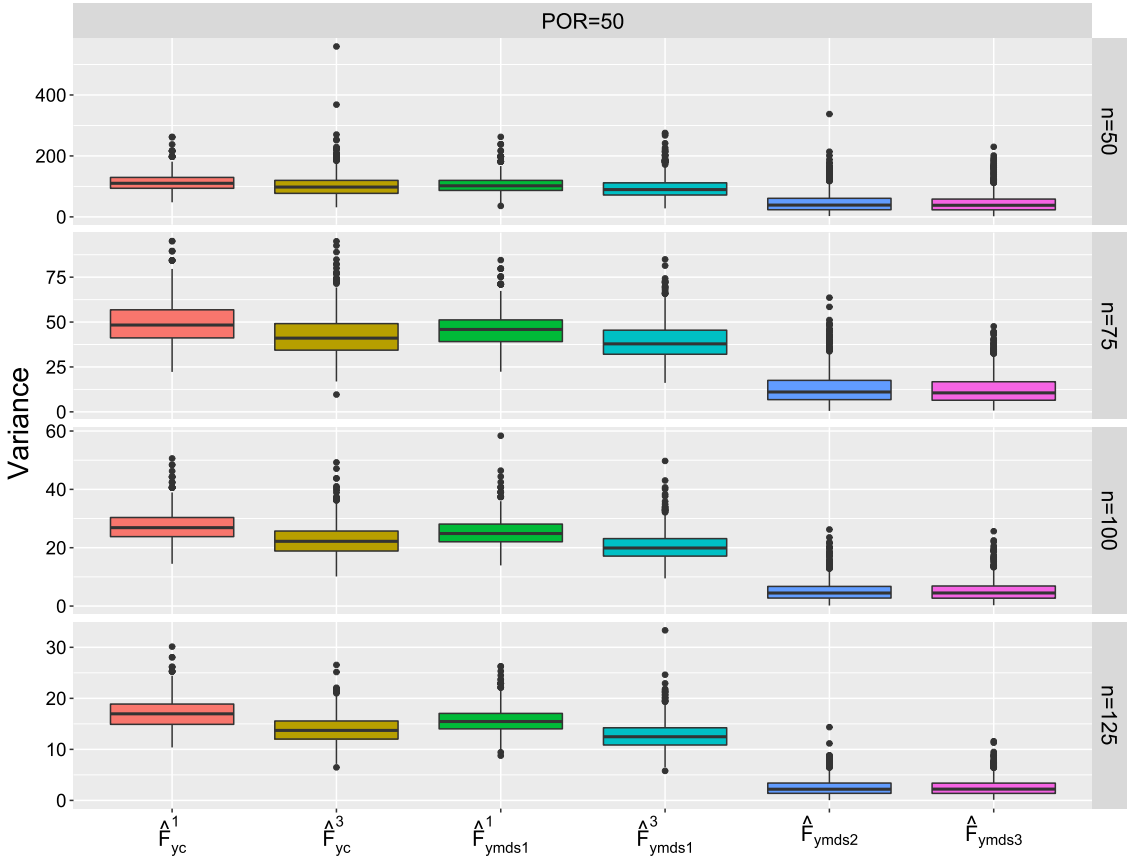

Fig. 5. Boxplot for variability of calibration weights SIMPOPULATION, $PE = 50$.

The proposed estimators present all the properties of a genuine distribution function under non-restrictive conditions. Perhaps, to satisfy the condition of nondecreasing monotony, we should ensure that the calibrated weights are nonnegative for the estimators $\widehat{F}_{ymds2}$ and $\widehat{F}_{ymds3}$ although this can also be achieved by using the raking distance in the calibration process.

In summary, we have conducted a simulation study with three different populations to compare the performance of the proposed methods with other indirect estimators of the distribution function. In this study, the distribution function was estimated with respect to a range of scenarios, with different sampling designs and estimations in the subpopulations. Additionally, we analysed the variability of the final set of calibrated weights of the proposed estimators, compared to the usual calibration estimators. From the results obtained, we conclude that the proposed estimators generally improve the relative efficiency of their respective

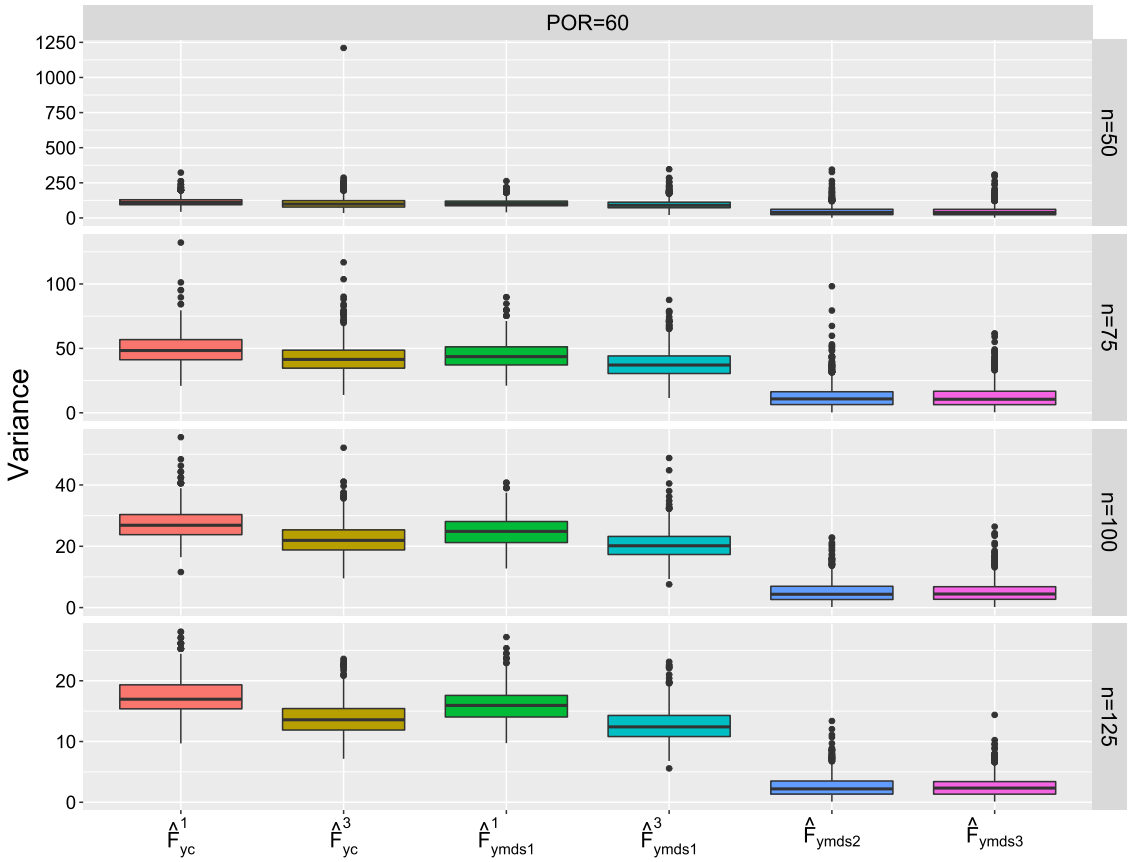

Fig. 6. Boxplot for variability of calibration weights SIMPOPULATION, $PE = 60$.

Table 25
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Women Subpopulation of SPANISH500, simple random sampling and $PE = 50\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{Y}}_{HT}$	0.0069	1.0000	0.0110	1.0000	0.0019	1.0000	0.0037	1.0000
$\widehat{\bar{Y}}_{CD}$	0.0156	0.5943	0.0174	0.5929	0.0130	0.5863	0.0138	0.6340
$\widehat{\bar{Y}}_{RKM}$	0.0072	0.6706	0.0052	0.6433	0.0051	0.6265	0.0029	0.6637
$\widehat{\bar{Y}}^1_{yc}$	0.0093	0.9602	0.0049	0.9477	0.0027	0.9560	0.0036	0.9756
$\widehat{\bar{Y}}^3_{yc}$	0.0085	0.8768	0.0063	0.8217	0.0043	0.8192	0.0027	0.8703
$\widehat{\bar{Y}}^1_{ymds1}$	0.0104	0.8554	0.0131	0.7876	0.0035	0.7402	0.0031	0.7467
$\widehat{\bar{Y}}^3_{ymds1}$	0.0091	0.7753	0.0141	0.6718	0.0082	0.6364	0.0048	0.6346
$\widehat{\bar{Y}}_{ymds2}$	0.0206	0.3928	0.0095	0.3437	0.0065	0.3352	0.0056	0.3555
$\widehat{\bar{Y}}_{ymds3}$	0.0231	0.4353	0.0102	0.3747	0.0071	0.3707	0.0056	0.3843

versions based on the usual calibration, and in some cases they also achieve a lower relative bias. In all the scenarios considered, one of the estimators that we propose always achieves the best efficiency compared to the other indirect estimators included for comparison purposes, and although no estimator consistently shows the best performance, $\widehat{\bar{Y}}_{ymds2}$ and $\widehat{\bar{Y}}_{ymds3}$ are generally the ones that significantly improve the efficiency of the estimates. However, in some cases, the estimator $\widehat{\bar{Y}}_{ymds3}$ presents efficiency problems and considerable variability in the final set of calibrated weights for high GOF values, perhaps derived from overcalibration problems. On the other hand, estimator $\widehat{\bar{Y}}_{ymds2}$ presents a more stable pattern of efficiency and its set of calibrated weights has the least variability of all the calibrated estimators. Consequently, we believe the estimator $\widehat{\bar{Y}}_{ymds2}$ is a reliable option for estimating the distribution function in the presence of auxiliary information that includes both qualitative and quantitative variables. For this reason and given that $\widehat{\bar{Y}}_{ymds2}$ can be used under any sampling design and as the calibrated weights neither depend on the study variable nor present excessive variability, we recommend their use in estimating the distribution function.

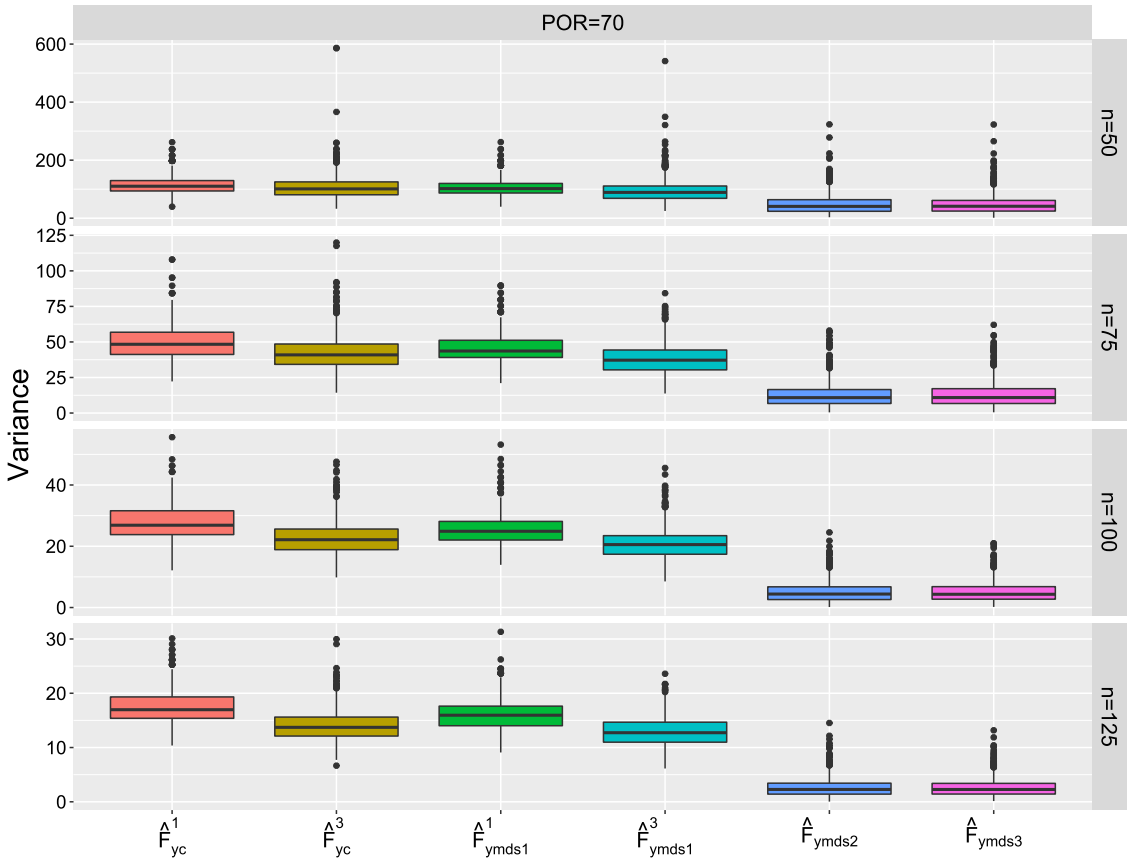

Fig. 7. Boxplot for variability of calibration weights SIMPOPULATION, $PE = 70$.

Table 26
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Women Subpopulation of SPANISH500, simple random sampling and $PE = 60\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\hat{\bar{Y}}_{HT}$	0.0040	1.0000	0.0045	1.0000	0.0039	1.0000	0.0058	1.0000
$\hat{\bar{Y}}_{CD}$	0.0189	0.6107	0.0160	0.6181	0.0129	0.6060	0.0115	0.5889
$\hat{\bar{Y}}_{RKM}$	0.0034	0.6815	0.0018	0.6637	0.0027	0.6439	0.0016	0.6217
$\hat{\bar{Y}}^1_{yc}$	0.0046	0.9340	0.0020	0.9410	0.0025	0.9331	0.0028	0.9222
$\hat{\bar{Y}}^3_{yc}$	0.0047	0.9247	0.0022	0.8648	0.0032	0.8349	0.0023	0.8003
$\hat{\bar{Y}}^1_{ymds1}$	0.0085	0.8413	0.0048	0.8080	0.0028	0.7755	0.0039	0.7547
$\hat{\bar{Y}}^3_{ymds1}$	0.0124	0.7570	0.0098	0.6921	0.0065	0.6460	0.0052	0.6123
$\hat{\bar{Y}}_{ymds2}$	0.0215	0.4084	0.0136	0.3585	0.0076	0.3399	0.0078	0.3347
$\hat{\bar{Y}}_{ymds3}$	0.0203	0.4365	0.0117	0.3831	0.0074	0.3705	0.0071	0.3620

Finally, this study is subject to certain limitations that could usefully be addressed in future research. Firstly, all the proposals we discuss are based on Gower's measure of similarity (1971). This is the most popular way of measuring the similarity/dissimilarity between observations in the presence of mixed variables, but some modifications of the unweighted distance have been proposed [48], seeking to balance the contribution of the different variables to the overall distance. Further analysis is needed to determine whether there exist other, more suitable, similarity measures for estimating the distribution function. Another question that remains to be considered is the optimal value of GOF taken to maximise the performance of the proposed distribution function estimators. Thirdly, our simulation study considers only simple random sampling or Midzuno sampling. It would be useful to examine how the proposed estimators behave in practice for other complex sample designs. Finally, additional research is needed to better characterise the performance of the proposed estimators when estimating population quantiles.

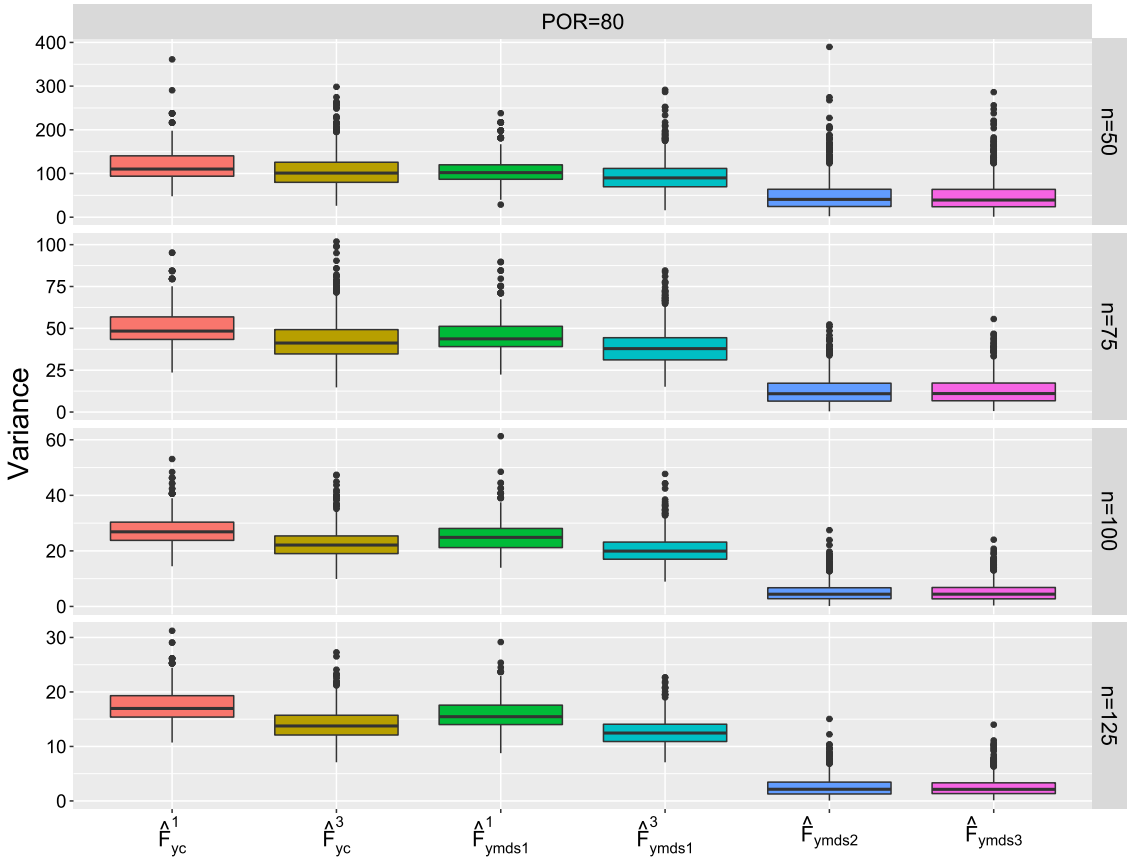

Fig. 8. Boxplot for variability of calibration weights SIMPOPULATION, $PE = 80$.

Table 27
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Women Subpopulation of SPANISH500, simple random sampling and $PE = 70\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{F}}_{HT}$	0.0084	1.0000	0.0050	1.0000	0.0036	1.0000	0.0063	1.0000
$\widehat{\bar{F}}_{CD}$	0.0221	0.6170	0.0193	0.6138	0.0162	0.6344	0.0137	0.6117
$\widehat{\bar{F}}_{RKM}$	0.0095	0.6881	0.0088	0.6705	0.0032	0.6781	0.0040	0.6441
$\widehat{\bar{F}}^1_{yc}$	0.0089	0.9784	0.0045	0.9482	0.0029	0.9816	0.0060	0.9458
$\widehat{\bar{F}}^3_{yc}$	0.0066	0.9143	0.0040	0.8679	0.0036	0.8679	0.0057	0.8510
$\widehat{\bar{F}}^1_{ymds1}$	0.0074	0.8382	0.0113	0.8162	0.0093	0.8660	0.0085	0.7665
$\widehat{\bar{F}}^3_{ymds1}$	0.0078	0.8014	0.0125	0.7111	0.0075	0.7413	0.0088	0.6442
$\widehat{\bar{F}}_{ymds2}$	0.0103	0.3495	0.0034	0.2944	0.0047	0.2942	0.0025	0.2611
$\widehat{\bar{F}}_{ymds3}$	0.0304	0.5479	0.0188	0.4481	0.0089	0.4159	0.0066	0.3720

Data availability

Data will be made available on request.

Acknowledgements

Funding

Grant PID2019-106861RB-I00 supported by MCIN/AEI/10.13039/501100011033, Spain.

Grant CEX2020-001105-M supported by Consejería de Universidad, Investigación e Innovación and by ERDF Andalusia Program 2021-2027, Andalusia, Spain.

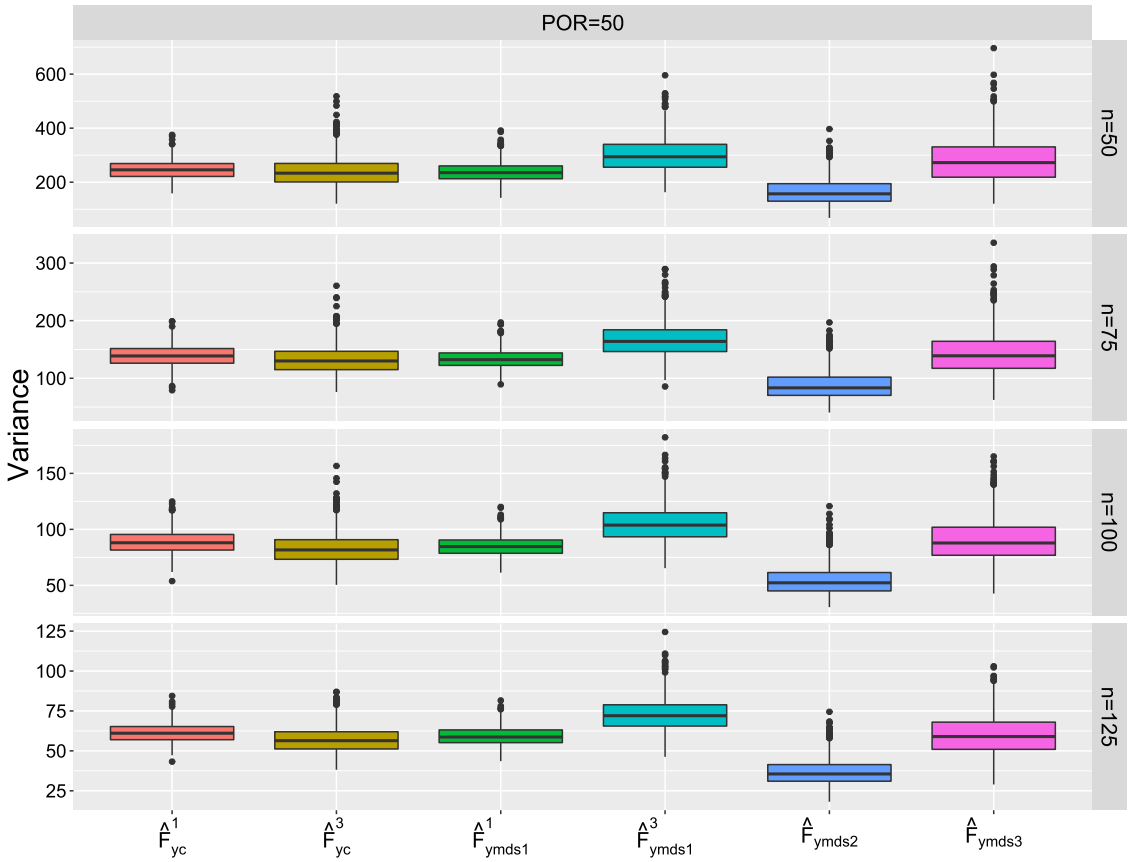

Fig. 9. Boxplot for variability of calibration weights EUSILC population, $PE = 50$.

Table 28
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Women Subpopulation of SPANISH500, simple random sampling and $PE = 80\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\hat{\bar{Y}}_{HT}$	0.0066	1.0000	0.0063	1.0000	0.0042	1.0000	0.0040	1.0000
$\hat{\bar{Y}}_{CD}$	0.0201	0.5641	0.0172	0.6039	0.0114	0.5930	0.0094	0.6157
$\hat{\bar{Y}}_{RKM}$	0.0054	0.6270	0.0057	0.6483	0.0029	0.6352	0.0046	0.6441
$\hat{\bar{Y}}^1_{yc}$	0.0097	0.9792	0.0081	0.9515	0.0027	0.9619	0.0059	0.9358
$\hat{\bar{Y}}^3_{yc}$	0.0072	0.8570	0.0063	0.8458	0.0022	0.8448	0.0055	0.8361
$\hat{\bar{Y}}^1_{ymds1}$	0.0177	0.4909	0.0096	0.5044	0.0040	0.4744	0.0049	0.5060
$\hat{\bar{Y}}^3_{ymds1}$	0.0130	0.2450	0.0065	0.2211	0.0028	0.2178	0.0037	0.2122
$\hat{\bar{Y}}_{ymds2}$	0.0086	0.4311	0.0084	0.3016	0.0078	0.2419	0.0067	0.2316
$\hat{\bar{Y}}_{ymds3}$	0.0471	1.3630	0.0260	0.6073	0.0130	0.4278	0.0114	0.4126

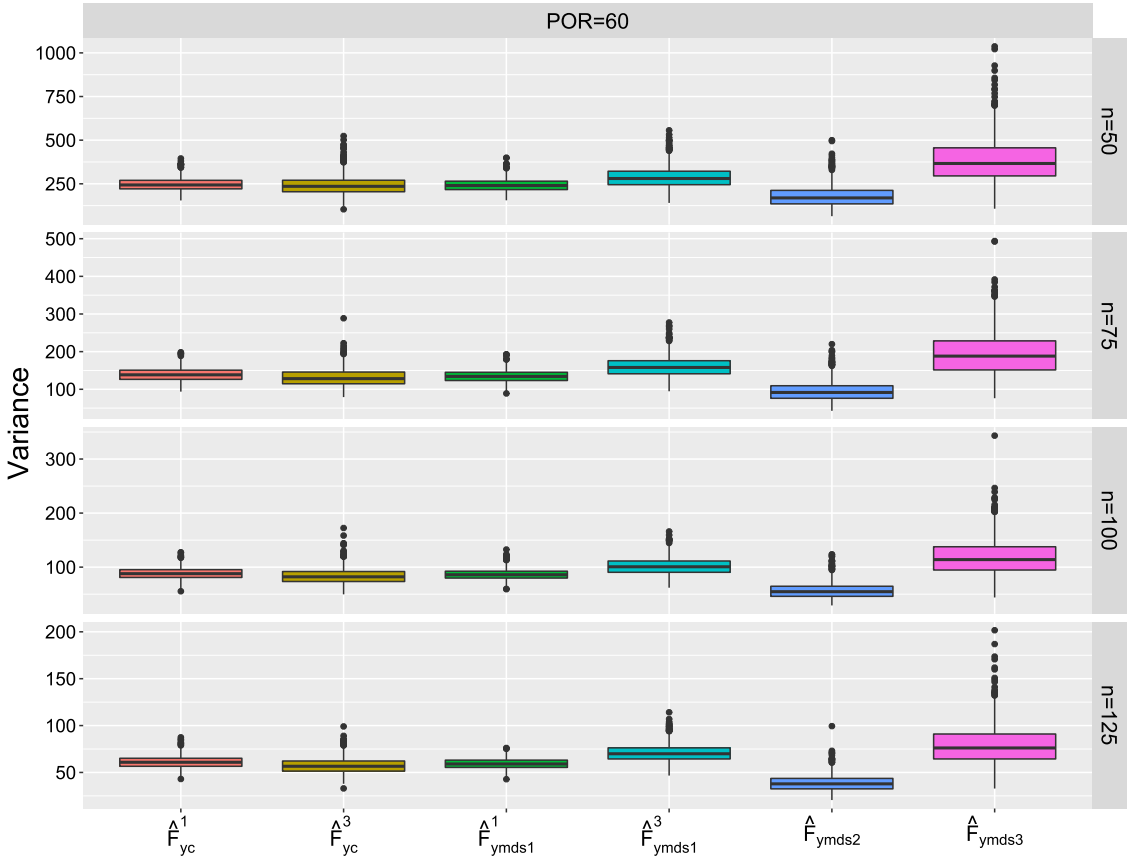

Fig. 10. Boxplot for variability of calibration weights EUSILC population, $PE = 60$.

Table 29
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. People with Internet access Subpopulation of SPANISH500, simple random sampling and $PE = 50\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{F}}_{HT}$	0.0076	1.0000	0.0089	1.0000	0.0034	1.0000	0.0030	1.0000
$\widehat{\bar{F}}_{CD}$	0.0196	0.6043	0.0099	0.6383	0.0105	0.6691	0.0135	0.6858
$\widehat{\bar{F}}_{RKM}$	0.0089	0.6858	0.0093	0.7091	0.0030	0.7227	0.0035	0.7322
$\widehat{\bar{F}}^{1}_{yc}$	0.0059	0.9994	0.0072	1.0065	0.0041	1.0146	0.0034	1.0639
$\widehat{\bar{F}}^{3}_{yc}$	0.0077	0.9473	0.0042	0.9323	0.0031	0.9200	0.0032	0.9419
$\widehat{\bar{F}}^{1}_{ymds1}$	0.0061	0.6798	0.0123	0.7020	0.0043	0.6841	0.0032	0.7131
$\widehat{\bar{F}}^{3}_{ymds1}$	0.0069	0.5554	0.0056	0.5437	0.0033	0.5281	0.0026	0.5370
$\widehat{\bar{F}}_{ymds2}$	0.0138	0.3217	0.0067	0.3320	0.0045	0.3055	0.0052	0.3278
$\widehat{\bar{F}}_{ymds3}$	0.0092	0.5528	0.0048	0.5589	0.0032	0.5164	0.0034	0.5501

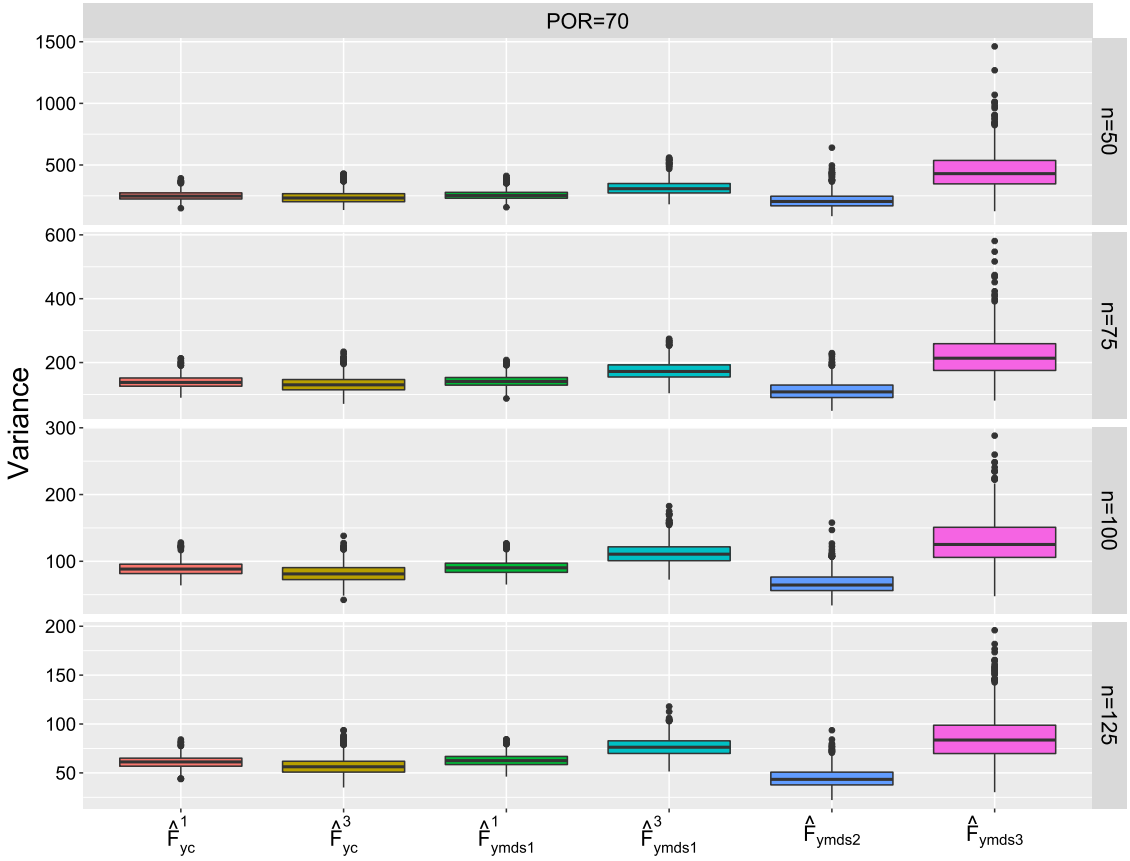

Fig. 11. Boxplot for variability of calibration weights EUSILC population, $PE = 70$.

Table 30
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. People with Internet access Subpopulation of SPANISH500, simple random sampling and $PE = 60\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{Y}}_{HT}$	0.0029	1.0000	0.0081	1.0000	0.0034	1.0000	0.0069	1.0000
$\widehat{\bar{Y}}_{CD}$	0.0142	0.6175	0.0131	0.5982	0.0133	0.6470	0.0089	0.6536
$\widehat{\bar{Y}}_{RKM}$	0.0041	0.6974	0.0052	0.6664	0.0031	0.7026	0.0052	0.6950
$\widehat{\bar{Y}}^{1}_{yc}$	0.0050	1.0088	0.0067	0.9686	0.0032	1.0026	0.0032	0.9730
$\widehat{\bar{Y}}^{3}_{yc}$	0.0066	0.9510	0.0056	0.9062	0.0035	0.8843	0.0037	0.8906
$\widehat{\bar{Y}}^{1}_{ymds1}$	0.0087	0.6998	0.0091	0.6802	0.0053	0.6442	0.0093	0.6652
$\widehat{\bar{Y}}^{3}_{ymds1}$	0.0040	0.5696	0.0063	0.5190	0.0033	0.5015	0.0041	0.4970
$\widehat{\bar{Y}}_{ymds2}$	0.0121	0.3405	0.0109	0.3254	0.0056	0.3231	0.0015	0.3171
$\widehat{\bar{Y}}_{ymds3}$	0.0069	0.5712	0.0075	0.5141	0.0045	0.5130	0.0029	0.5152

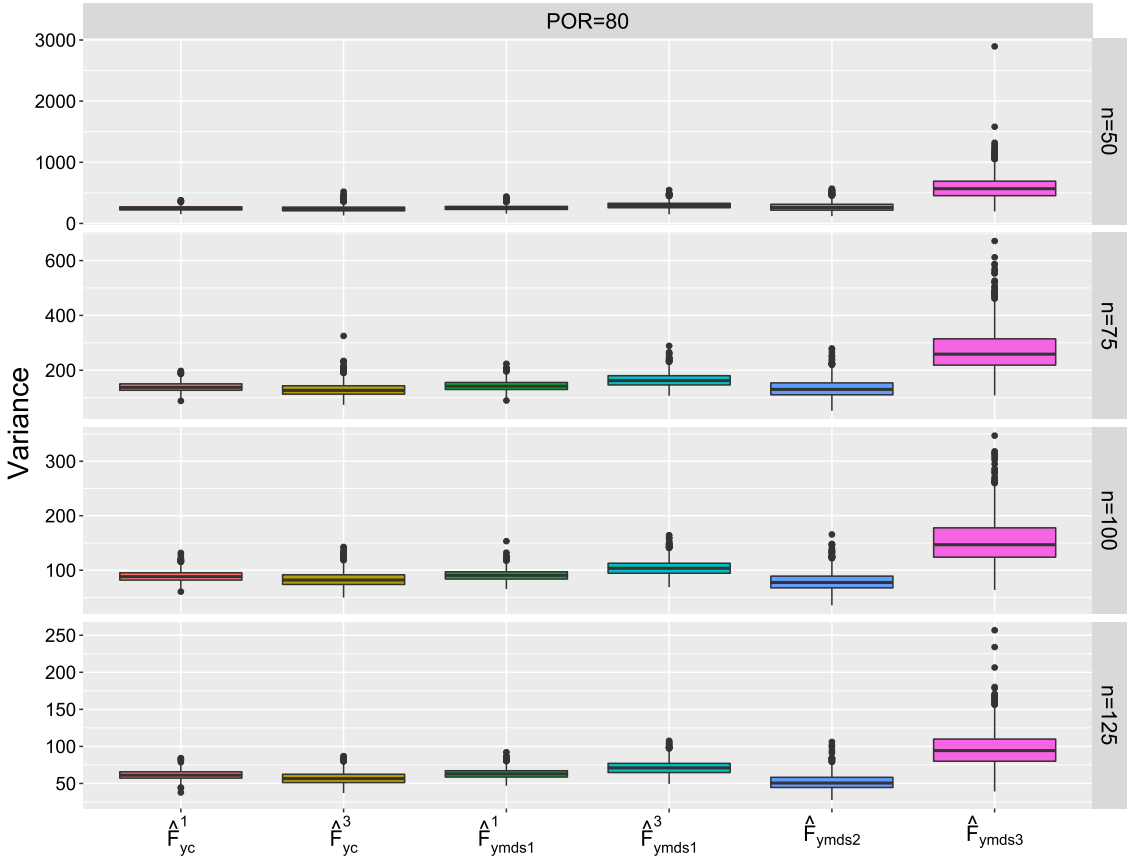

Fig. 12. Boxplot for variability of calibration weights EUSILC population, $PE = 80$.

Table 31
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. People with Internet access Subpopulation of SPANISH500, simple random sampling and $PE = 70\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{Y}}_{HT}$	0.0036	1.0000	0.0035	1.0000	0.0040	1.0000	0.0034	1.0000
$\widehat{\bar{Y}}_{CD}$	0.0148	0.6252	0.0105	0.6554	0.0102	0.6450	0.0139	0.6371
$\widehat{\bar{Y}}_{RKM}$	0.0029	0.7075	0.0042	0.7173	0.0028	0.7061	0.0047	0.6825
$\widehat{\bar{Y}}^1_{yc}$	0.0039	0.9903	0.0038	0.9850	0.0031	1.0082	0.0039	0.9911
$\widehat{\bar{Y}}^3_{yc}$	0.0060	0.9129	0.0037	0.8755	0.0025	0.9050	0.0043	0.8888
$\widehat{\bar{Y}}^1_{ymds1}$	0.0082	0.7246	0.0110	0.6780	0.0040	0.6908	0.0037	0.6675
$\widehat{\bar{Y}}^3_{ymds1}$	0.0022	0.6215	0.0077	0.5423	0.0027	0.5487	0.0025	0.5062
$\widehat{\bar{Y}}_{ymds2}$	0.0069	0.2375	0.0033	0.2076	0.0034	0.2057	0.0036	0.1868
$\widehat{\bar{Y}}_{ymds3}$	0.0151	0.6625	0.0059	0.5677	0.0092	0.5694	0.0090	0.5330

Table 32
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. People with Internet access Subpopulation of SPANISH500, simple random sampling and $PE = 80\%$.

	$n = 50$		$n = 75$		$n = 100$		$n = 125$	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
$\widehat{\bar{F}}_{HT}$	0.0024	1.0000	0.0029	1.0000	0.0037	1.0000	0.0014	1.0000
$\widehat{\bar{F}}_{CD}$	0.0133	0.6340	0.0123	0.6139	0.0141	0.6734	0.0127	0.6332
$\widehat{\bar{F}}_{RKM}$	0.0024	0.7182	0.0030	0.6827	0.0032	0.7236	0.0015	0.6760
$\widehat{\bar{F}}^{1}_{yc}$	0.0076	1.0145	0.0036	1.0228	0.0051	0.9884	0.0018	1.0045
$\widehat{\bar{F}}^{3}_{yc}$	0.0097	0.9665	0.0023	0.9213	0.0041	0.9180	0.0035	0.8633
$\widehat{\bar{F}}^{1}_{ymds1}$	0.0030	0.5571	0.0043	0.5345	0.0014	0.5170	0.0016	0.4922
$\widehat{\bar{F}}^{3}_{ymds1}$	0.0099	0.2356	0.0071	0.2202	0.0026	0.2029	0.0033	0.2111
$\widehat{\bar{F}}_{ymds2}$	0.0072	0.2375	0.0060	0.1724	0.0029	0.1467	0.0022	0.1490
$\widehat{\bar{F}}_{ymds3}$	0.1570	1.6472	0.1152	1.4035	0.0738	1.1705	0.0535	1.1398

References

[1] N. Sedransk, J. Sedransk, Distinguishing among distributions using data from complex sample designs, J. Amer. Statist. Assoc. 74 (368) (1979) 754–760, http://dx.doi.org/10.1080/01621459.1979.10481028.

[2] Ç. Çetinkaya, A.I. Genç, Stress–strength reliability estimation under the standard two-sided power distribution, Appl. Appl. Math. Model. 65 (2019) 72–88, http://dx.doi.org/10.1016/j.apm.2018.08.008.

[3] J.F. Muñoz, P.J. Moya-Fernández, E. Álvarez-Verdejo, Exploring and correcting the bias in the estimation of the gini measure of inequality, Sociol. Methods Res. (2023) http://dx.doi.org/10.1177/00491241231176847.

[4] E. Álvarez-Verdejo, P.J. Moya-Fernández, J.F. Muñoz-Rosas, Single imputation methods and confidence intervals for the gini index, Mathematics 9 (2021) 3252, http://dx.doi.org/10.3390/math9243252.

[5] S. Martínez, M. Illescas, H. Martínez, A. Arcos, Calibration estimator for head count index, Int. J. Comput. Math. 97 (1–2) (2020) 51–62, http://dx.doi.org/10.1080/00207160.2018.1425798.

[6] J. Foster, J. Greer, E. Thorbecke, A class of decomposable poverty measures, Econometrica 52 (3) (1984) 761–766, http://dx.doi.org/10.2307/1913475.

[7] S. Martínez, M. Rueda, M. Illescas, The optimization problem of quantile and poverty measures estimation based on calibration, J. Comput. Appl. Math. 405 (2022) 113054, http://dx.doi.org/10.1016/j.cam.2020.113054.

[8] M. Rueda, S. Martínez-Puertas, H. Martínez-Puertas, A. Arcos, Calibration methods for estimating quantiles, Metrika 66 (3) (2007) 355–371, http://dx.doi.org/10.1007/s00184-006-0116-1.

[9] S. Martínez, M. Rueda, A. Arcos, H. Martínez, I. Sánchez-Borrego, Post-stratified calibration method for estimating quantiles, Comput. Statist. Data Anal. 55 (1) (2011) 838–851, http://dx.doi.org/10.1016/j.csda.2010.07.006.

[10] M.K. Bohn, V. Higgins, P. Kavsak, B. Hoffman, K. Adeli, High-sensitivity generation 5 cardiac troponin t sex-and age-specific 99th percentiles in the CALIPER cohort of healthy children and adolescents, Clin. Chem. 65 (4) (2019) 589–591, http://dx.doi.org/10.1373/clinchem.2018.299156.

[11] S.T. Wolford, R.A. Schroer, F.X. Gohs, P.P. Gallo, M. Brodeck, H.B. Falk, R. Ruhren, Reference range data base for serum chemistry and hematology values in laboratory animals, J. Toxicol. Environ. Health A 18 (2) (1986) 161–188.

[12] R. Bu, J. Lu, T. Ren, B. Liu, X. Li, R. Cong, Particulate organic matter affects soil nitrogen mineralization under two crop rotation systems, PLoS One 10 (12) (2015) e0143835, http://dx.doi.org/10.1371/journal.pone.0143835.

[13] R. Decker, J. Haltiwanger, R. Jarmin, J. Miranda, The role of entrepreneurship in US job creation and economic dynamism, J. Econ. Perspect. 28 (3) (2014) 3–24, http://dx.doi.org/10.1257/jep.28.3.3.

[14] G. Chauvet, C. Goga, Asymptotic efficiency of the calibration estimator in a high-dimensional data setting, J. Statist. Plann. Inference 217 (2022) 177–187, http://dx.doi.org/10.1016/j.jspi.2021.07.011.

[15] J.C. Deville, C.E. Särndal, Calibration estimators in survey sampling, J. Amer. Statist. Assoc. 87 (418) (1992) 376–382, http://dx.doi.org/10.1080/01621459.1992.10475217.

[16] A. Arcos, S. Martínez, M. Rueda, H. Martínez, Distribution function estimates from dual frame context, J. Comput. Appl. Math. 318 (2017) 242–252, http://dx.doi.org/10.1016/j.cam.2016.09.027.

[17] T. Harms, P. Duchesne, On calibration estimation for quantiles, Surv. Methodol. 32 (2006) 37–52.

[18] J.A. Mayor-Gallego, J.L. Moreno-Rebollo, M.D. Jiménez-Gamero, Estimation of the finite population distribution function using a global penalized calibration method, AStA Adv. Stat. Anal. 103 (1) (2019) 1–35, http://dx.doi.org/10.1007/s10182-018-0321-z.

[19] M. Rueda, S. Martínez, H. Martínez, A. Arcos, Estimation of the distribution function with calibration methods, J. Statist. Plann. Inference 137 (2) (2007) 435–448, http://dx.doi.org/10.1016/j.jspi.2005.12.011.

[20] H.P. Singh, S. Singh, M. Kozak, A family of estimators of finite-population distribution function using auxiliary information, Acta Appl. Math. 104 (2) (2008) 115–130, http://dx.doi.org/10.1007/s10440-008-9243-1.

[21] C. Wu, Optimal calibration estimators in survey sampling, Biometrika 90 (4) (2003) 937–951, http://dx.doi.org/10.1093/biomet/90.4.937.

[22] S. Martínez, M. Rueda, A. Arcos, H. Martínez, Optimum calibration points estimating distribution functions, J. Comput. Appl. Math. 233 (9) (2010) 2265–2277, http://dx.doi.org/10.1016/j.cam.2009.10.011.

[23] S. Martínez, M. Rueda, A. Arcos, H. Martínez, J.F. Muñoz, On determining the calibration equations to construct model-calibration estimators of the distribution function, Rev. Mat. Complut. 25 (1) (2012) 87–95, http://dx.doi.org/10.1007/s13163-010-0058-z.

[24] S. Martínez, M. Rueda, H. Martínez, A. Arcos, Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function, J. Comput. Appl. Math. 318 (2017) 444–459, http://dx.doi.org/10.1016/j.cam.2016.02.002.

[25] J.F. Vera, C.C. Sánchez Zuleta, M.D.M. Rueda, A unified approach based on multidimensional scaling for calibration estimation in survey sampling with qualitative auxiliary information, Stat. Methods Med. Res. (2023) http://dx.doi.org/10.1177/09622802231151211.

[26] D. Devaud, Y. Tillé, Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem, Test 28 (4) (2019) 1033–1065, http://dx.doi.org/10.1007/s11749-019-00681-3.

[27] C.E. Särndal, The calibration approach in survey theory and practice, Surv. Methodol. 33 (2) (2007) 99–119.

[28] M. Rueda, S. Martinez, H. Martinez, A. Arcos, Mean estimation with calibration techniques in presence of missing data, Comput. Statist. Data Anal. 50 (11) (2006) 3263–3277, http://dx.doi.org/10.1016/j.csda.2005.06.003.

[29] M. Rueda, S. Martinez, A. Arcos, J.F. Muñoz, Mean estimation under successive sampling with calibration estimators, Comm. Statist. Theory Methods 38 (6) (2009) 808–827, http://dx.doi.org/10.1080/03610920802316609.
[30] M.G. Ranalli, A. Arcos, M.M. Rueda, A. Teodoro, Calibration estimation in dual-frame surveys, Stat. Methods Appl. 25 (3) (2016) 321–349, http://dx.doi.org/10.1007/s10260-015-0336-5.
[31] P.L.D. Nascimento Silva, C.J. Skinner, Variable selection for regression estimation in finite populations, Surv. Methodol. 23 (1) (1997) 23–32.
[32] J.N.K. Rao, A.C. Singh, Range restricted weight calibration for survey data using ridge regression, Pak. J. Stat. 25 (4) (2009).
[33] F. Guggemos, Y. Tille, Penalized calibration in survey sampling: Design-based estimation assisted by mixed models, J. Statist. Plann. Inference 140 (11) (2010) 3199–3212, http://dx.doi.org/10.1016/j.jspi.2010.04.010.
[34] H. Cardot, C. Goga, M.A. Shehzad, Calibration and partial calibration on principal components when the number of auxiliary variables is large, Statist. Sinica (2017) 243–260, https://www.jstor.org/stable/44114370.
[35] R.G. Clark, R.L. Chambers, Adaptive calibration for prediction of finite population totals, Surv. Methodol. 34 (2008) 163–172, 172.
[36] J.F. Beaumont, C. Bocci, Another look at ridge calibration, Metron 66 (2008) 5–20.
[37] S. Martínez, M. Rueda, M.D. Illescas, Reduction of optimal calibration dimension with a new optimal auxiliary vector for calibrated estimators of the distribution function, Math. Methods Appl. Sci. 45 (17) (2022) 10959–10981, http://dx.doi.org/10.1002/mma.8431.
[38] S. Martínez, M.D. Illescas, M. Rueda, Distribution function estimation with calibration on principal components, J. Comput. Appl. Math. 428 (2023) 115189, http://dx.doi.org/10.1016/j.cam.2023.115189.
[39] S. Martínez, M. Rueda, H. Martínez, A. Arcos, Determining P optimum calibration points to construct calibration estimators of the distribution function, J. Comput. Appl. Math. 275 (2015) 281–293, http://dx.doi.org/10.1016/j.cam.2014.07.020.
[40] J.C. Gower, A general coefficient of similarity and some of its properties, Biometrics 27 (4) (1971) 857–871, http://dx.doi.org/10.2307/2528823.
[41] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, New York, 1979.
[42] M. Illescas-Manzano, S. Martínez-Puertas, M.d. Mar Rueda, A. Arcos-Rueda, Calibration adjustment for dealing with nonresponse in the estimation of poverty measures, in: D. Barrera, S. Remogna, D. Sbibih (Eds.), Mathematical and Computational Methods for Modelling, Approximation and Simulation, in: SEMA SIMAI Springer Series, vol. 29, Springer, Cham, 2022, http://dx.doi.org/10.1007/978-3-030-94339-4_11.
[43] J.C. Deville, C.E. Särndal, O. Sautory, Generalized raking procedures in survey sampling, J. Amer. Statist. Assoc. 88 (423) (1993) 1013–1020.
[44] R.L. Chambers, R. Dunstan, Estimating distribution functions from survey data, Biometrika 73 (3) (1986) 597–604, http://dx.doi.org/10.1093/biomet/73.3.597.
[45] J.N.K. Rao, J.G. Kovar, H.J. Mantel, On estimating distribution functions and quantiles from survey data using auxiliary information, Biometrika 77 (2) (1990) 365–375, http://dx.doi.org/10.2307/2336815.
[46] F.J. Breidt, J.D. Opsomer, Local polynomial regression estimators in survey sampling, Ann. Statist. 28 (4) (2000) 1026–1053, https://www.jstor.org/stable/2673953.
[47] I. Borg, PJF. Groenen, Modern Multidimensional Scaling. Theory and Applications, second ed., Springer, New York, 2005.
[48] M. D'Orazio, Distances with mixed type variables some modified Gower's coefficients, 2021, ArXiv, abs/2101.02481.