Journal of Computational and Applied Mathematics 428 (2023) 115189

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Distribution function estimation with calibration on principal components

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. University of Granada, Avenida de la Fuente Nueva S/N, Granada, 18071, Spain

ARTICLE INFO

Article history:
Received 5 September 2022
Received in revised form 24 February 2023

MSC:
62D05

Keywords:
Auxiliary information
Distribution function
Calibration technique
Principal components
Survey sampling

ABSTRACT

The calibration method is a convenient means of incorporating auxiliary information when several parameters must be estimated. This approach has recently been used to develop new estimators for the distribution function. However, the auxiliary information available may generate a large dataset, provoking a loss of efficiency in the estimators obtained, due to over-calibration. We propose adapting the calibration using principal components, in order to avoid the negative consequences of over-calibration when estimating the distribution function.

1. Introduction

In sampling surveys, estimation of the distribution function is an important issue in many current areas of research. By estimating the distribution function, researchers can then use it to evaluate parameters such as quantiles [1], which are of great interest in many areas [2–4]. For example, in economics, this parameter enables the analyst to estimate measures of poverty [5,6]) and inequality [1,7,8]. Indeed, in some cases knowledge of the distribution function is more useful than calculating total and mean values [9].

With current technological advances, very large sets of auxiliary variables are commonly generated [10]. To facilitate their incorporation at the estimation stage, some authors have proposed alternative estimators of the distribution function [11–16] by means of calibration [17]. One such proposal [14] provides a computationally simple way to incorporate the auxiliary information and produces estimators that are genuine distribution functions under mild conditions. However, the asymptotic behaviour of the proposed method depends on the choice of an auxiliary vector [18,19] and although both the optimal dimension of this auxiliary vector and its optimal choice in order to minimise the variance have been established [20], the optimal dimension can have a large value, even in the reduced version developed by [21].

The high dimension of the vector that optimises the asymptotic behaviour of the proposed estimator by [14] can be an important issue. Firstly, the calibration process may not have a solution. Furthermore, a large dimension of auxiliary information can cause over-calibration, which would increase the bias of the calibration estimators and can reduce their efficiency [10]. To avoid these problems, the previous research has proposed solutions in the estimation of totals or

* 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.2023.115189
0377-0427/

means [22–25], such as procedures for variable selection [22] or penalised calibration [23,24], an approach that takes auxiliary information into account in various ways, by attaching greater or lesser importance to this information according to its presumed explanatory power for the variable of interest [24]. Under this approach, [13] proposed a new model-assisted estimator of the distribution function that combined ideas underlying model calibration and penalised calibration. In another approach, [25] proposed performing the calibration on the R first principal components, thus avoiding the problems derived from over-calibration in the estimation of totals and means.

In this paper, our aim is to build upon the method proposed by [14] and by [20] for estimating the distribution function, by adapting the calibration with principal components described by [25]. In Section 2, we present the context of our proposal and discuss classical estimation procedures for the distribution function. Section 3 then introduces four new estimators of the distribution function, derived from the joint application of the methods proposed by [14,25]. The properties of these calibrated estimators are described in Section 4, after which Section 5 presents the results obtained from a simulation study. The paper concludes in Section 6 with some final remarks.

2. Estimation of the distribution function and calibration estimation

Let $U = \{1, \ldots, N\}$ be a finite population with size N and let $s = \{1, 2, \ldots, n\}$ be a random sample of fixed size n, drawn from U with a specified sampling design $p(\cdot)$ that assigns known inclusion probabilities of first and second order denoted by π_k and π_{kl}, $k, l \in U$, respectively. The value $d_k = \pi_k^{-1}$ denotes the corresponding sampling design weight for unit $k \in U$. Let Y be the variable of interest and let y_k be the value of Y for unit k. We assume that y_k is known for all sample units. Our aim is to estimate the distribution function $F_y(t)$ for the study variable Y which can be defined as follows

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

where $\Delta(\cdot)$ denotes the Heaviside function, given by

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

Without auxiliary information, the distribution function $F_y(t)$ can be unbiasedly estimated by the Horvitz–Thompson estimator, defined by

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

Now, if we consider J auxiliary variables $X_1, \ldots, X_J$ and let $\mathbf{x}_k' = (x_{1k}, \ldots, x_{Jk})$ be the vector of auxiliary variables at unit k, we assume that $\mathbf{x}_k$ is available for all population units (complete auxiliary information). Although the estimator $\widehat{F}_{YHT}(t)$ is unbiased, it does not incorporate the auxiliary information provided by the auxiliary vector $\mathbf{x}_k$.

The calibration method, originally developed by Deville and Särndall [17] for the estimation of totals or means, enables us to incorporate the auxiliary information available through the auxiliary vector $\mathbf{x}_k$. This method was adapted by [14] to estimate the distribution function $F_y(t)$, from the definition of the following pseudo-variable

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

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

The calibration method could now be applied from the pseudo-variable g. To do so, the basic weights d_k are replaced by new weights ω_k, minimising the chi-square distance

$$\Phi_s = \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k}, \tag{5}$$

subject to the following conditions

$$\frac{1}{N} \sum_{k \in s} \omega_k \Delta(t_j - g_k) = F_g(t_j) \quad j = 1, 2, \ldots, P. \tag{6}$$

where $F_g(t_j)$ is the distribution function of g at the points t_j, $j = 1, 2, \ldots, P$, where, with no loss in generality, $t_1 < t_2 < \cdots t_P$. The values q_k are positive constants unrelated to d_k. Assuming that the matrix T given by

$$\sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k) \Delta(\mathbf{t_g} - g_k)',$$

is nonsingular, the resulting calibration estimator is

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

with

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

and $\widehat{F}_{GHT}(\mathbf{t_g})$ denoting the Horvitz–Thompson estimator of $F_g(\mathbf{t_g})$ at $\mathbf{t_g} = (t_1, \ldots, t_P)'$.

Following [14], the main advantages of $\widehat{F}_{yc}(t)$ are

- $\widehat{F}_{yc}(t)$ is a genuine distribution function and therefore it can be applied directly to estimate quantiles.
- $\widehat{F}_{yc}(t)$ is asymptotically unbiased.
- The asymptotic variance is known and is given by

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

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

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

Among the disadvantages of $\widehat{F}_{yc}(t)$

- The calibration procedure to obtain $\widehat{F}_{yc}(t)$ is based on a pseudo-variable g. Therefore $\widehat{F}_{yc}(t)$ assumes that the auxiliary variables and the study variable are linearly related.
- The asymptotic behaviour of $\widehat{F}_{yc}(t)$ and its precision depends on the choice of auxiliary vector $\mathbf{t_g}$

Under simple random sampling, the optimal choice of the vector $\mathbf{t_g}$ has recently been analysed [20,26,27] in order to minimise the asymptotic variance (8). In this respect, [20] proposed a calibration estimator for $F_y(t)$ based on the optimal dimension for $\mathbf{t_g}$ and its optimal selection $\mathbf{t_{opt}}$.

3. Distribution function estimation based on principal components

To incorporate the calibration with principal components into the estimator proposed by [14,20], we will consider three alternatives

1. Firstly, if the auxiliary vector $\mathbf{x}_k'$ has a large dimension, following the method proposed in [25] for the estimation of totals or means, the dimension of $\mathbf{x}_k'$ can be reduced by employing principal components. Thus, we obtain a new auxiliary vector $\mathbf{C}_k' = (C_{1k}, \ldots, C_{Rk})$ with the first R components associated with the auxiliary vector $\mathbf{x}_k'$ where $R < J$, which provides the calibrated estimator of the distribution function as proposed by [14] from this new vector $\mathbf{C}_k'$.
2. Secondly, under simple random sampling, we can consider the calibration estimator for the distribution function $F_y(t)$ proposed by [14] based on the optimal auxiliary vector from [20] or its reduced version as in [21], but instead of calibrating with the optimal auxiliary vector $\mathbf{t_{opt}}$, which can produce a large dimension, we can reduce its dimension by obtaining the principal components of the set of auxiliary variables that appear in the condition (6) associated with $\mathbf{t_{opt}}$.
3. Finally, for all the variables x_{ik} included in the auxiliary vector $\mathbf{x}_k'$ we can consider the following vector with dimension N

$$(\mathbf{t}_k^i)' = \left(\Delta(x_{i1} - x_{ik}), \ldots, \Delta(x_{iN} - x_{ik})\right), \quad k \in U, \quad i = 1, \ldots, J.$$

Let us now consider the final auxiliary vector with dimension $N \cdot J$

$$\mathbf{T}_k' = ((\mathbf{t}_k^1)', \ldots, (\mathbf{t}_k^J)'), \quad k \in U. \tag{10}$$

With the auxiliary vector $\mathbf{T}_k$, we can access all the information related to the distribution functions of the variables included in the auxiliary vector $\mathbf{x}_k$. However, the calibration process based directly on the auxiliary vector $\mathbf{T}_k$ cannot be solved due to the large number of restrictions imposed, many of which may be incompatible. Therefore, reducing the dimension of the vector $\mathbf{T}_k$ by means of principal components allows us to make best use of the auxiliary information. This way, we can reduce the dimension of $\mathbf{T}_k'$ to obtain a new, reduced auxiliary vector $\mathbf{Z}_k' = (Z_{1k}, \ldots, Z_{Lk})$ with $L < N \cdot J$ and then build a new calibration estimator for $F_y(t)$ based on the new variables.

3.1. Distribution function estimation based on principal components of the auxiliary vector $\mathbf{x}'_k$

If the auxiliary vector $\mathbf{x}'_k$ has a large dimension, there may be multicollinearity among the auxiliary variables. In such a case, the estimator (4) is sensitive to small changes in $\mathbf{x}_k$ and y_k and has a large variance [25]. To avoid this issue, [25] proposed reducing the dimension of $\mathbf{x}'_k$ when means or totals must be estimated. This issue may also affect the efficiency of the estimator proposed by [14] for the distribution function $F_y(t)$ due to the pseudo-variable g. Thus, we can obtain the first R principal components $\mathbf{C}'_k = (C_{1k}, \ldots, C_{Rk})$ and build a new pseudo-variable a in the following way

$$a_k = \widehat{\beta}'_C \mathbf{C}_k \text{ for } k = 1, 2, \ldots, N,$$

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

By minimising (5) subject to the constraints

$$\frac{1}{N} \sum_{k \in s} \omega_k \Delta(v_j - a_k) = F_a(v_j), \quad j = 1, 2, \ldots, P, \tag{12}$$

where $F_a(v_j)$ is the distribution function of a at the points v_j, $\quad j = 1, 2, \ldots, P$ where $v_1 < v_2 < \cdots v_P$.

The calibration estimator obtained for $F_y(t)$ is given by

$$\widehat{F}_{ycomp1}(t) = \widehat{F}_{YHT}(t) + \left(F_a(\mathbf{v_a}) - \widehat{F}_{AHT}(\mathbf{v_a}) \right)' \cdot \widehat{D}(\mathbf{v_a}), \tag{13}$$

assuming that the matrix H given by

$$H = \sum_{k \in s} d_k q_k \Delta(\mathbf{v_g} - a_k) \Delta(\mathbf{v_a} - a_k)',$$

is nonsingular,

$$\widehat{D}(\mathbf{v_a}) = H^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\mathbf{v_a} - a_k) \Delta(t - y_k),$$

and $\widehat{F}_{AHT}(\mathbf{t_g})$ is the Horvitz–Thompson estimator of $F_a(\mathbf{v_a})$ at $\mathbf{v_a} = (v_1, \ldots, v_P)'$.

Following [14], $\widehat{F}_{yc1}(t)$ is asymptotically unbiased and the asymptotic variance is

$$AV(\widehat{F}_{ycomp1}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (d_k U_k)(d_l U_l), \tag{14}$$

where $U_k = \Delta(t - y_k) - \Delta(\mathbf{v_a} - a_k)' \cdot D(\mathbf{v_a})$, with

$$D(\mathbf{v_a}) = \left(\sum_{k \in U} q_k \Delta(\mathbf{v_a} - a_k) \Delta(\mathbf{v_a} - a_k)' \right)^{-1} \cdot \left(\sum_{k \in U} q_k \Delta(\mathbf{v_a} - a_k) \Delta(t - y_k) \right). \tag{15}$$

Additionally, if $q_k = c$ and v_P is large enough, the estimator $F_{yc1}(t)$ is a genuine distribution function [14].

3.2. Distribution function estimation based on principal components of the optimal auxiliary vector $\mathbf{t_{opt}}$

Now, if the sample s is obtained by simple random sampling and if $q_k = c$ for all $k \in U$, we can apply the optimal auxiliary vector [20] or its reduced version [21] in order to minimise asymptotic variance of the calibration estimator $\widehat{F}_{yc}(t)$ given by (8). Since the estimator $\widehat{F}_{yc}(t)$ depends on the choice of the vector $\mathbf{t}_g$, [20] calculated the optimal dimension of $\mathbf{t_g}$ and the optimal vector $\mathbf{t}_{opt}$ for a value t from the following sets

$$A_t = \{g_k : k \in U; y_k \leq t\} = \{a_1^t, a_2^t, \ldots, a_{M_t}^t\} \quad \text{with} \quad a_h^t < a_{h+1}^t, \tag{16}$$

for $h = 1, \ldots, M_t - 1$, where M_t is the number of elements in the set A_t and

$$B_t = \{b_1^t, b_2^t, \ldots, b_{M_t}^t\},$$

with

$$b_1^t = \max_{l \in U_1} \{g_l\} \quad \text{where } U_1 = \{l \in U : g_l < a_1^t\},$$
$$b_h^t = \max_{l \in U_h} \{g_l\} \quad \text{where } U_h = \{l \in U : a_{h-1}^t \leq g_l < a_h^t\}, \quad h = 2, 3, \ldots, M_t,$$

and $b_h^t \leq b_{h+1}^t$ for $h = 1, \ldots, M_t - 1$.

Thus, following [20] the optimal dimension for the auxiliary vector $\mathbf{t}_g$ is $P = 2M_t$ if b_1^t exists and for $j = 2, \ldots, M_t$, $b_j^t \neq a_{j-1}^t$ and the optimal auxiliary vector $\mathbf{t}_{opt}$ is given by

$$\mathbf{t_{opt}}(t) = (b_1^t, a_1^t, \ldots, b_{M_t}^t, a_{M_t}^t). \tag{17}$$

If there are values $j_1^t, j_2^t, \ldots, j_{p_t}^t \in \{1, \ldots, M_t\}$ for which $a_{j_1-1}^t = b_{j_1}^t$ with $p_t \leq M_t$ and $j_h^t \neq j_q^t$ if $h \neq q$, then the optimal dimension is $P = 2M_t - p_t$ and the optimal auxiliary vector $\mathbf{t_{opt}}$ is

$$\mathbf{t_{opt}}(t) = (b_1^t, a_1^t \ldots, b_{j_h-1}^t, a_{j_h-1}^t, a_{j_h}^t, b_{j_h+1}^t, \ldots, b_{M_t}^t, a_{M_t}^t). \tag{18}$$

Since $\mathbf{t}_{opt}$ depends on the population values of the study variable Y, we must consider a sample version of $\mathbf{t}_{opt}$. To avoid the over-calibration problems associated with a high dimension of the optimal auxiliary vector, instead of calibrating with $\mathbf{t}_{opt} = (t_{1opt}, \ldots, t_{Mopt})$, we can consider the set of auxiliary variables associated with $\mathbf{t_{opt}}$ given by

$$\mathbf{\Delta}_k^{opt}(t) = \big(\Delta(t_{1opt} - g_k), \ldots, \Delta(t_{Mopt} - g_k)\big), \quad k \in U.$$

If the dimension of the vector $\mathbf{\Delta}_k^{opt}(t)$ is large, it can be reduced by considering the first $R(t)$ components $\mathbf{C}_k(t)' = (C_{1k}(t), \ldots, C_{R(t)k}(t))$. Let us now consider the calibration estimator obtained by minimising (5) subject to the following conditions

$$\frac{1}{N}\sum_{k\in s}\omega_k\mathbf{C}_k(t) = \frac{1}{N}\sum_{k\in U}\mathbf{C}_k(t) = \bar{\mathbf{C}}. \tag{19}$$

The calibration estimator obtained for $F_y(t)$ is given by

$$\widehat{F}_{ycomp2}(t) = \widehat{F}_{YHT}(t) + \left(\bar{\mathbf{C}} - \widehat{\bar{C}}_{HT}\right)' \cdot \widehat{D}_C, \tag{20}$$

assuming that the matrix H_C given by

$$H_C = \sum_{k\in s} d_k q_k \mathbf{C}_k(t) \cdot \mathbf{C}_k(t)',$$

is nonsingular,

$$\widehat{D}(C) = H_C^{-1} \cdot \sum_{k\in s} d_k q_k \mathbf{C}_k(t) \Delta(t - y_k),$$

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

Additionally, we can consider the reduced version of $\mathbf{t}_{opt}$ proposed in [21], i.e. $\mathbf{t}_{optred}$. In the same way, we can propose a second version estimator $\widehat{F}_{ycomp2red}(t)$ based on the principal components of the auxiliary variables related to $\mathbf{t}_{optred}$.

3.3. Distribution function estimation based on principal components of the auxiliary vector $\mathbf{T}_k'$

Finally, under a general sampling design, we can consider the auxiliary vector $\mathbf{T}_k'$ given by (10) with dimension $N \cdot J$. The dimension of $\mathbf{T}_k'$ can be reduced by principal components to obtain a new reduced auxiliary vector $\mathbf{Z}_k' = (Z_{1k}, \ldots, Z_{Lk})$ with $L < N \cdot J$ and then build a new calibration estimator for $F_y(t)$ based on the new variables by minimising (5) under the constraints:

$$\frac{1}{N}\sum_{k\in s}\omega_k\mathbf{Z}_k = \frac{1}{N}\sum_{k\in U}\mathbf{Z}_k = \bar{\mathbf{Z}}. \tag{21}$$

The new calibration estimator is given by

$$\widehat{F}_{ycomp3}(t) = \widehat{F}_{YHT}(t) + \left(\bar{\mathbf{Z}} - \widehat{\bar{Z}}_{HT}\right)' \cdot \widehat{D}_Z, \tag{22}$$

assuming that the matrix H_Z given by

$$H_Z = \sum_{k\in s} d_k q_k \mathbf{Z}_k \cdot \mathbf{Z}_k',$$

is nonsingular,

$$\widehat{D}(Z) = H_Z^{-1} \cdot \sum_{k\in s} d_k q_k \mathbf{Z}_k \Delta(t - y_k),$$

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

Following [14], $\widehat{F}_{ycomp3}(t)$ is asymptotically unbiased and the asymptotic variance is

$$AV(\widehat{F}_{ycomp3}(t)) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k W_k)(d_l W_l), \tag{23}$$

where $W_k = \Delta(t - y_k) - \mathbf{Z}_k' \cdot D(\mathbf{Z})$, with

$$D(\mathbf{Z}) = \left(\sum_{k\in U} q_k \mathbf{Z}_k \mathbf{Z}_k'\right)^{-1} \cdot \left(\sum_{k\in U} q_k \mathbf{Z}_k \Delta(t - y_k)\right). \tag{24}$$

3.4. Principal components selection criteria

A central issue in principal component analysis is how to determine the number of principal components to retain. In this subsection, we consider some commonly used criteria, such as the scree plot and the proportion of variance extracted, together with alternatives that avoid the use of negative calibrated weights.

Without loss of generality, we assume that the auxiliary variables included in the auxiliary vector $\mathbf{x}_k$ are centred (i.e. that the variables have a zero mean). Let $\mathbf{X}$ be the $N \times J$ matrix with $\mathbf{x}_k$, $k \in U$ as rows. The variance–covariance matrix is then given by $N^{-1}\mathbf{X}^T \cdot \mathbf{X}$. Let $C_1, C_2, \ldots, C_J$ be the J principal components associated with $\mathbf{X}$ and let $\lambda_1 \geq \lambda_2 \geq \cdots > \lambda_J$ be the eigenvalues of the variance–covariance matrix $N^{-1}\mathbf{X}^T \cdot \mathbf{X}$, such that

$$Var(C_j) = \lambda_j, \quad \text{for } j = 1, 2 \ldots J,$$

and

$$\sum_{j=1}^{J} Var(x_j) = \sum_{j=1}^{J} Var(C_j) = \sum_{j=1}^{J} \lambda_j.$$

Some of the most widely considered methods and criteria are [28]

- Percentage of total variance explained.
- Kaiser–Guttman criterion.
- Scree plot criterion.

Regarding the first of these criteria, the percentage of the total variance represented by the first R principal components is given by

$$PV = \frac{\sum_{j=1}^{R} \lambda_j}{\sum_{j=1}^{J} \lambda_j} \cdot 100. \tag{25}$$

In this procedure, components are selected until a certain percentage of total variance p_0 is covered, such as $p_0 = 70\%$, $p_0 = 80\%$ or $p_0 = 90\%$. In other words, the first R components are retained so that $PV \geq p_0$ [29].

With the Kaiser–Guttman criterion [28], the components retained are those with an eigenvalue that is greater than the average of eigenvalues, i.e., the components for which the eigenvalue verifies

$$\lambda_j > \bar{\lambda} = \frac{\sum_{j=1}^{J} \lambda_j}{J}.$$

Since variables are often measured in different units, the use of the correlation matrix is considered in the analysis of principal components in order to standardise the variables. In this case $\bar{\lambda} = 1$ and this criterion selects the components C_j with the eigenvalue $\lambda_j > 1$.

Another widely used criterion is the scree plot [28], in which the value of each successive eigenvalue λ_j is plotted against j. The smaller eigenvalues lie along a straight line, that is, there is a point on the scree plot from which the eigenvalues are approximately equal. The criterion retains those components whose eigenvalue does not fall on this line.

Finally, since calibrated weights based on principal components do not have to be positive for all sample units in general, [25] proposed a criterion to select the number R of principal components based on the tuning parameter selection, in a ridge regression context suggested in [30] that avoids negative calibrated weights. Specifically, under the selection strategy for the principal components dimension suggested in [25], the dimension R plays the role of a tuning parameter. The largest dimension R is then chosen such that all the calibration weights ω_k remain positive for all sample units (for further details of this criterion see [25,30]).

4. Properties of the calibrated estimators based on principal components

In estimating the distribution function, it is important to know whether a proposed estimator satisfies the distribution function properties, that is, whether the new estimator is a genuine distribution function. For an estimator $\widehat{F}_y(t)$ of $F_y(t)$ to be a genuine distribution function, the following conditions must be satisfied

(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.

If the estimator $\widehat{F}_y(t)$ satisfies all these conditions, we can estimate quantiles and hence achieve appropriate poverty measures based on quantiles by considering the inverse of $\widehat{F}_y(t)$ [31]. It should be noted that the calibrated estimators proposed in the previous section may not present all the properties of the distribution function.

Clearly, the proposed estimators $\widehat{F}_{ycomp1}(t)$, $\widehat{F}_{ycomp2}(t)$, $\widehat{F}_{ycomp2red}(t)$ and $\widehat{F}_{ycomp3}(t)$ satisfy conditions i) and iia), but in general they do not fulfil conditions iib) and iii).

Regarding the first of these estimators $\widehat{F}_{ycomp1}(t)$, following [14] condition iib) can be met if a sufficiently large value of v_P is selected in vector $\mathbf{v_a}$. Moreover, condition iii) is satisfied if we choose $q_k = c$ for all units in the population. Henceforth, we assume these conditions when calculating the estimator $\widehat{F}_{ycomp1}(t)$.

Concerning estimators $\widehat{F}_{ycomp2}(t)$, $\widehat{F}_{ycomp2red}(t)$ and $\widehat{F}_{ycomp3}(t)$, condition iib) is met if the calibration weights ω_k satisfy the following constraint

$$\frac{1}{N}\sum_{k\in s}\omega_k = 1. \tag{26}$$

Restriction (26) can be included in the calibration process for each estimator $\widehat{F}_{ycomp2}(t)$, $\widehat{F}_{ycomp2red}(t)$ and $\widehat{F}_{ycomp3}(t)$. Therefore, we consider both constraint (26) and the corresponding constraints (19) and (21) in order to satisfy condition iib). Henceforth, it is assumed that constraint (26) is included in the calibration processes of each of the estimators $\widehat{F}_{ycomp2}(t)$, $\widehat{F}_{ycomp2red}(t)$ and $\widehat{F}_{ycomp3}(t)$.

Finally, verifying the non-decreasing monotonicity of the estimators $\widehat{F}_{ycomp2}(t)$, $\widehat{F}_{ycomp2red}(t)$ and $\widehat{F}_{ycomp3}(t)$ is equivalent to verifying that the calibrated weights are positive, that is, $\omega_k \geq 0$ for all $k \in s$. As mentioned above, calibrated weights based on principal components do not have to be positive for all sample units, so in general estimators $\widehat{F}_{ycomp2}(t)$, $\widehat{F}_{ycomp2red}(t)$ and $\widehat{F}_{ycomp3}(t)$ do not meet condition iii). In order to satisfy the condition of non-decreasing monotonicity, under the criterion proposed by [25] the optimum principal components dimension can be selected by choosing the largest dimensions $R(t)$ and L such that all the calibration weights ω_k remain positive.

5. Simulation study

To analyse the behaviour of the proposed estimators, we carried out a simulation study using new routines programmed in R [version 4.2.1]. The performance of these estimators $\widehat{F}_{ycomp1}(t)$, $\widehat{F}_{ycomp2}(t)$, $\widehat{F}_{ycomp2red}(t)$ and $\widehat{F}_{ycomp3}(t)$ was compared with that of other estimators of the distribution function $F_y(t)$. Specifically, the new estimators were compared with the Horvitz–Thompson estimator, $\widehat{F}_{HT}(t)$, the difference estimator [32], $\widehat{F}_D(t)$, the ratio estimator [32] $\widehat{F}_R(t)$, the Chambers–Dunstan estimator [33] $\widehat{F}_{CD}(t)$ and the Kovar–Mantel–Rao-Estimator [32].

Regarding the calibration estimator $\widehat{F}_{yc}(t)$ [14], four alternatives were considered. For the first one $\widehat{F}^1_{yc}(t)$, the vector selected was $\mathbf{t_g} = (Q_g(0.5))$. For the second one $\widehat{F}^3_{yc}(t)$ was taken as $\mathbf{t_g} = (Q_g(0.25), Q_g(0.5), Q_g(0.75))$. In the third alternative $\widehat{F}_{ycopt}(t)$, we took $\mathbf{tg} = \mathbf{t_{opt}}$ and in the last alternative $\widehat{F}_{ycoptred}(t)$, we assumed $\mathbf{tg} = \mathbf{t_{optred}}$.

Finally, regarding the proposed estimator $\widehat{F}_{ycomp1}(t)$, two versions were included. The first one $\widehat{F}^1_{ycomp1}(t)$ was based on $\mathbf{v_a} = (Q_a(0.5))$ and the second $\widehat{F}^3_{ycomp1}(t)$ on $\mathbf{v_a} = (Q_a(0.25), Q_a(0.5), Q_a(0.75))$.

The first population considered is a small real population, termed SUGAR CANE, which was previously studied by [32,33]. This population consists of 338 sugar cane farms in Queensland, Australia, that were surveyed in 1982 and characterised by means of four variables: total cane harvested, gross value of the cane, total farm expenditure and area dedicated to the crop. The study variable is the total farm expenditure and the remaining variables are included in the auxiliary vector $\mathbf{x}_k$. Although this auxiliary vector is very small, the condition number of $N^{-1}\mathbf{X}^T \cdot \mathbf{X}$ is 47504.25, where the matrix $\mathbf{X}$ has the auxiliary vector $\mathbf{x}'_k$ as rows. Therefore, there are strong correlations among the variables used in the calibration process.

The second population, SIMPOPULATION, is a generated population of size $N = 1000$, with 16 variables based on the procedure described in [34]. The values for the first variable η_k were generated as independent and identically distributed (i.i.d.) from a uniform distribution in (0, 1). Additionally, the following variables were generated, based on different regression models

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

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

In this case, the study variable is c_{2k} and the remaining variables are included in the auxiliary vector $\mathbf{x}_k$, so that the condition number of $N^{-1}\mathbf{X}^T \cdot \mathbf{X}$ was 260975828.

The third population considered was the EUSILC dataset included in the R package "laeken". This database was synthetically generated from real data published in EU Statistics on Income and Living Conditions in Austria. The study variable was employee cash or near cash income (net), and the remaining variables were included in the auxiliary vector $\mathbf{x}_k$. For the qualitative variables, corresponding dummy variables were considered. The final auxiliary vector dimension was $J = 38$. As some data were missing, only complete units for all the variables were considered in the simulation study, producing a total population size of $N = 12107$. The condition number of $N^{-1}\mathbf{X}^T \cdot \mathbf{X}$ was 111060589.

For each population, four different sizes were considered. One thousand samples were drawn for each size by simple random sampling without replacement. For each sample, estimations of the distribution function $F(t)$ were obtained by each of the estimators included in the simulation study, at 11 different points, namely the quantiles $Q_y(\alpha)$ for α=0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8 and 0.9. Since our aim in this simulation study was to estimate $F_y(t)$ instead of obtaining the quantile estimation, we did not apply the selection criteria of the principal components dimension to meet condition iii) as discussed above. Instead, the criterion applied was the percentage of total variance explained, based on the *PV* given by (25). Thus, different values of *PV*, $PV = 60\%, 70\%, 80\%$ and 90%, were considered in each population.

The performance of each estimator included in the simulation study was determined by calculating the average relative bias (AVRB) and the average relative efficiency (AVRE) in each case. These measures are given by

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

where

$$\text{RB}(t) = \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{F}(t)_b - F_y(t)}{F_y(t)} \quad \text{and} \quad \text{RE}(t) = \frac{MSE[\widehat{F}(t)]}{MSE[\widehat{F}_{HT}(t)]}, \tag{27}$$

where b indexes the bth simulation run, $\widehat{F}(t)$ is an estimator of the distribution function, $MSE[\widehat{F}(t)] = B^{-1}\sum_{b=1}^{B}[\widehat{F}(t)_b - F_y(t)]^2$ is the empirical mean square error for $\widehat{F}(t)$ and $MSE[\widehat{F}_{HT}(t)]$ is similarly defined for the Horvitz–Thompson estimator.

Tables 1–3 respectively provide the results for the SUGAR CANE, SIMPOPULATION and EUSILC populations. For the SUGAR CANE population, all of the estimators give similar values for bias (AVRB), except the $\widehat{F}_{CD}$, $\widehat{F}_{ycopt}$, $\widehat{F}_{yoptred}$, $\widehat{F}_{ycomp2}$ and $\widehat{F}_{ycomp2red}$, which present a greater degree of bias. The versions based on principal components ($\widehat{F}_{ycomp2}$ and $\widehat{F}_{ycomp2red}$) have lower values than the respective versions ($\widehat{F}_{ycopt}$ and $\widehat{F}_{yoptred}$) without calibration based on principal components.

The best results for efficiency (AVRE) are produced by the estimators $\widehat{F}_{ycomp2}$; $\widehat{F}_{ycomp2red}$ and $\widehat{F}_{ycomp3}$, followed by $\widehat{F}_{CD}$ and $\widehat{F}^3_{yc}$. For the percentage of variance extracted $PV = 60\%$ and $PV = 70\%$, the most efficient estimator is $\widehat{F}_{ycomp2red}$. For $PV = 80\%$, the most efficient estimators are $\widehat{F}_{ycomp2}$ and $\widehat{F}_{ycomp3}$. For $PV = 90\%$, the best estimators are $\widehat{F}_{ycomp2}$ and $\widehat{F}_{ycomp2red}$, while $\widehat{F}_{ycomp3}$ is inefficient compared to $\widehat{F}_{HT}$ for sample sizes $n = 50$ and $n = 75$ and again shows better efficiency than most estimators for sample sizes $n = 100$ and $n = 125$. In general, the estimator $\widehat{F}^1_{ycomp1}$, although it is more efficient than $\widehat{F}_{HT}$, presents the worst efficiency among the estimators that incorporate auxiliary information, while $\widehat{F}^3_{ycomp1}$ presents worse efficiency than $\widehat{F}^3_{yc}$ (concerning the version not based on principal components).

Our analysis of SIMPOPULATION (Table 2) shows that all of the estimators produce similar results for bias, with the possible exceptions of $\widehat{F}_R$, $\widehat{F}_{CD}$, $\widehat{F}_{ycopt}$ and $\widehat{F}_{ycoptred}$, which show higher values for sample sizes $n = 50$; $n = 75$ and in some cases for $n = 100$. Regarding efficiency, the best estimators are clearly $\widehat{F}^3_{ycomp1}$ and $\widehat{F}_{ycomp3}$. While $\widehat{F}^1_{ycomp1}$ presents greater efficiency than the other estimators, its value is very close to that of $\widehat{F}_{HT}$ for $PV = 60\%$. As the value of *PV* increases, the efficiency of $\widehat{F}^1_{ycomp1}$ increases. The estimator $\widehat{F}_{ycomp3}$, as in the previous population, for $PV = 90\%$ is less efficient than $\widehat{F}_{HT}$ for $n = 50$ and $n = 75$ while for the other sample sizes, $\widehat{F}_{ycomp3}$ performs better. The remaining indirect estimators are less efficient than $\widehat{F}_{HT}$, except for $\widehat{F}_{CD}$ and $\widehat{F}_D$. The poor efficiency of most indirect estimators may be due to the high condition number of $N^{-1}\mathbf{X}^T \cdot \mathbf{X}$ and to the fact that a greater number of auxiliary variables are available than is the case with the SUGAR CANE population.

Finally, regarding the EUSILC population (Table 3), again most of the estimators present similar results for bias. The lowest values are obtained by $\widehat{F}_{ycomp2}$ and $\widehat{F}_{ycomp2red}$, except for $PV = 60\%$ and $n = 12$ where the estimator $\widehat{F}_{ycomp3}$ has the lowest bias. For $PV = 90\%$, this same estimator is the worst for bias. Similarly, with efficiency the best estimators

Table 1
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population: SUGAR CANE.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$PV = 60\%$								$PV = 70\%$							
	$n = 50$		$n = 75$		$n = 100$		$n = 125$		$n = 50$		$n = 75$		$n = 100$		$n = 125$	
$\widehat{F}_{HT}$	0.0039	1	0.0022	1	0.0013	1	0.0077	1	0.0034	1	0.0052	1	0.0021	1	0.0035	1
$\widehat{F}_{D}$	0.0039	1	0.0022	1	0.0013	1	0.0077	1	0.0034	1	0.0052	1	0.0021	1	0.0035	1
$\widehat{F}_{R}$	0.0039	1	0.0022	1	0.0013	1	0.0077	1	0.0034	1	0.0052	1	0.0021	1	0.0035	1
$\widehat{F}_{CD}$	0.0910	0.4376	0.0878	0.5237	0.0806	0.6201	0.0713	0.6589	0.0911	0.4515	0.0887	0.5371	0.0780	0.5894	0.0728	0.6910
$\widehat{F}_{RKM}$	0.0035	0.5342	0.0030	0.5396	0.0014	0.5522	0.0031	0.5408	0.0032	0.5468	0.0049	0.5449	0.0025	0.5210	0.0021	0.5521
$\widehat{F}^{1}_{yc}$	0.0044	0.8447	0.0023	0.9053	0.0011	0.8996	0.0045	0.8619	0.0033	0.8782	0.0048	0.8678	0.0015	0.8979	0.0033	0.8870
$\widehat{F}^{3}_{yc}$	0.0034	0.4636	0.0023	0.4687	0.0005	0.4547	0.0035	0.4543	0.0047	0.4562	0.0043	0.4587	0.0016	0.4624	0.0017	0.4550
$\widehat{F}_{ycopt}$	0.0492	0.5568	0.0296	0.5226	0.0227	0.5313	0.0151	0.5011	0.0533	0.5747	0.0341	0.5442	0.0207	0.5040	0.0162	0.5027
$\widehat{F}_{ycoptred}$	0.0492	0.5568	0.0296	0.5226	0.0227	0.5313	0.0151	0.5011	0.0533	0.5747	0.0341	0.5442	0.0207	0.5040	0.0162	0.5027
$\widehat{F}^{1}_{ycomp1}$	0.0048	0.8920	0.0022	0.9627	0.0017	0.9539	0.0045	0.9187	0.0032	0.9138	0.0052	0.9429	0.0022	0.9323	0.0033	0.9155
$\widehat{F}^{3}_{ycomp1}$	0.0028	0.5571	0.0019	0.5577	0.0016	0.5580	0.0035	0.5416	0.0038	0.5602	0.0032	0.5633	0.0013	0.5460	0.0012	0.5404
$\widehat{F}_{ycomp2}$	0.0235	0.4313	0.0125	0.4094	0.0105	0.4148	0.0078	0.3933	0.0223	0.4060	0.0133	0.3796	0.0082	0.3559	0.0061	0.3393
$\widehat{F}_{ycomp2red}$	0.0301	0.3561	0.0154	0.3019	0.0122	0.3079	0.0089	0.2858	0.0314	0.3597	0.0174	0.3261	0.0113	0.2914	0.0076	0.2762
$\widehat{F}_{ycomp3}$	0.0036	0.3856	0.0038	0.3768	0.0011	0.3820	0.0031	0.3734	0.0035	0.3677	0.0022	0.3381	0.0027	0.3214	0.0018	0.3192
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$PV = 80\%$								$PV = 90\%$							
	$n = 50$		$n = 75$		$n = 100$		$n = 125$		$n = 50$		$n = 75$		$n = 100$		$n = 125$	
$\widehat{F}_{HT}$	0.0069	1	0.0026	1	0.0028	1	0.0043	1	0.0045	1	0.0030	1	0.0014	1	0.0008	1
$\widehat{F}_{D}$	0.0049	0.6175	0.0038	0.6064	0.0020	0.6542	0.0043	1	0.0045	1	0.0030	1	0.0014	1	0.0008	1
$\widehat{F}_{R}$	0.0411	1.1608	0.0423	1.1773	0.0200	1.0875	0.0043	1	0.0045	1	0.0030	1	0.0014	1	0.0008	1
$\widehat{F}_{CD}$	0.0916	0.4457	0.0877	0.5084	0.0792	0.6362	0.0709	0.6649	0.0906	0.4576	0.0873	0.5448	0.0788	0.6247	0.0714	0.6679
$\widehat{F}_{RKM}$	0.0057	0.5406	0.0030	0.5289	0.0026	0.5764	0.0029	0.5439	0.0039	0.5653	0.0012	0.5518	0.0016	0.5499	0.0009	0.5409
$\widehat{F}^{1}_{yc}$	0.0054	0.8983	0.0021	0.9099	0.0027	0.9014	0.0048	0.8993	0.0053	0.9253	0.0019	0.9053	0.0014	0.8637	0.0012	0.9074
$\widehat{F}^{3}_{yc}$	0.0085	0.4631	0.0022	0.4775	0.0026	0.4813	0.0017	0.4681	0.0041	0.4842	0.0026	0.4720	0.0028	0.4654	0.0014	0.4491
$\widehat{F}_{ycopt}$	0.0546	0.5858	0.0334	0.5158	0.0224	0.5437	0.0163	0.4934	0.0549	0.6186	0.0327	0.5460	0.0225	0.5191	0.0159	0.5155
$\widehat{F}_{ycoptred}$	0.0546	0.5858	0.0334	0.5158	0.0224	0.5437	0.0163	0.4934	0.0549	0.6186	0.0327	0.5460	0.0225	0.5191	0.0159	0.5155
$\widehat{F}^{1}_{ycomp1}$	0.0051	0.9043	0.0019	0.9293	0.0029	0.9288	0.0044	0.9612	0.0047	0.9632	0.0021	0.9625	0.0014	0.9300	0.0007	0.9650
$\widehat{F}^{3}_{ycomp1}$	0.0054	0.5579	0.0028	0.5686	0.0031	0.5861	0.0031	0.5649	0.0045	0.5783	0.0019	0.5605	0.0017	0.5614	0.0010	0.5514
$\widehat{F}_{ycomp2}$	0.0316	0.3714	0.0182	0.3272	0.0131	0.3291	0.0089	0.2956	0.0304	0.3675	0.0184	0.3137	0.0130	0.2941	0.0085	0.2819
$\widehat{F}_{ycomp2red}$	0.0292	0.4420	0.0175	0.3582	0.0125	0.3257	0.0087	0.2789	0.0341	0.4167	0.0194	0.3377	0.0127	0.3050	0.0081	0.2812
$\widehat{F}_{ycomp3}$	0.0047	0.3767	0.0046	0.3134	0.0035	0.2957	0.0016	0.2775	0.0262	1.8936	0.0072	1.6652	0.0049	0.4310	0.0024	0.3552

are $\widehat{F}^{3}_{ycomp1}$, $\widehat{F}_{ycomp3}$ and $\widehat{F}^{1}_{ycomp1}$. The remaining indirect estimators are inefficient compared to $\widehat{F}_{HT}$, except for $\widehat{F}_{D}$, which in some cases is very slightly more efficient than $\widehat{F}_{HT}$. In this population, as in the previous case, we also have a high number of variables and a high condition number, which could explain the poor efficiency of most indirect estimators. Again, for $PV = 90\%$, the efficiency of $\widehat{F}_{ycomp3}$ worsens and becomes less efficient than $\widehat{F}_{HT}$, but in this case it takes place for all sample sizes.

6. Conclusions

In recent years, the calibration technique has made it possible to develop new estimators for finite populations that allow the available auxiliary information to be incorporated. This advance has attracted significant attention in survey sampling research and survey applications. However, calibrated estimators can suffer from over-calibration [10], that is, the loss of efficiency when high-dimensional auxiliary information is available, which with current technological advances is a very common situation in survey applications. Our study extends the calibration techniques developed by [25] based on principal components to the estimation of the distribution function of a variable of interest when complete auxiliary information is available. To the best of our knowledge, this aspect of the question has not been addressed previously.

In this paper, we propose three different methods for using principal components to perform calibration, from which we can estimate the distribution function. The estimators based on method 1 (calibrating on the principal components of auxiliary vector $\mathbf{x}'_k$) and those based on method 3 (calibrating on principal components of the auxiliary vector $\mathbf{T}'_k$) are formulated under a general sampling design, while the estimators based on method 2 (calibrating on the principal components of auxiliary vector $\mathbf{t_{opt}}$) are more restrictive in that they can only be used for simple random sampling. Another advantage of the estimators based on methods 1 and 3 is that the principal components are obtained on a vector of variables that does not depend on the objective variable, which is very useful in multipurpose surveys in which there are many main variables. On the contrary, with method 2 we must first obtain the vector $\mathbf{t_{opt}}$, which depends on the variable y, and then calculate the principal components for the vector associated with each main variable. In consequence, this method is more complex numerically.

Estimators based on any of these three methods have good properties: for example, they are continuous on the right, $\lim_{t\to-\infty}\widehat{F}_y(t) = 0$ and $\lim_{t\to+\infty}\widehat{F}_y(t) = 1$, if the restriction that the mean of the calibrated weights is 1 is added. To guarantee the non-decreasing monotonicity of the estimators based on methods 2 and 3, a particular selection strategy is needed for the principal components dimension.

Table 2
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population: SIMPOPULATION.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$PV = 60\%$								$PV = 70\%$							
	$n = 50$		$n = 75$		$n = 100$		$n = 125$		$n = 50$		$n = 75$		$n = 100$		$n = 125$	
$\widehat{F}_{HT}$	0.0016	1	0.0042	1	0.0033	1	0.0043	1	0.0054	1	0.0038	1	0.0026	1	0.0015	1
$\widehat{F}_{D}$	0.0016	1	0.0042	1	0.0033	1	0.0043	1	0.0054	1	0.0038	1	0.0026	1	0.0015	1
$\widehat{F}_{R}$	0.0216	2.2241	0.0138	2.2423	0.0063	2.0962	0.0043	1	0.0125	2.2458	0.0174	2.1530	0.0091	2.1377	0.0015	1
$\widehat{F}_{CD}$	0.0101	0.9538	0.0070	0.9404	0.0072	0.9534	0.0080	0.9585	0.0101	0.9520	0.0113	0.9682	0.0071	0.9481	0.0058	0.9546
$\widehat{F}_{RKM}$	0.0026	1.0149	0.0039	0.9999	0.0033	0.9983	0.0049	1.0075	0.0060	1.0201	0.0043	1.0144	0.0027	1.0037	0.0015	1.0055
$\widehat{F}^{1}_{yc}$	0.0053	1.4445	0.0007	1.4057	0.0019	1.4757	0.0056	1.4925	0.0064	1.5059	0.0042	1.4737	0.0021	1.3904	0.0012	1.4524
$\widehat{F}^{3}_{yc}$	0.0018	1.3790	0.0046	1.3431	0.0039	1.3336	0.0040	1.3240	0.0064	1.3575	0.0038	1.3449	0.0023	1.3003	0.0016	1.3492
$\widehat{F}_{ycopt}$	0.0150	2.0338	0.0119	1.9369	0.0116	2.0090	0.0070	2.0639	0.0241	1.9011	0.0176	1.9751	0.0082	1.9266	0.0084	1.9676
$\widehat{F}_{ycoptred}$	0.0150	2.0338	0.0119	1.9369	0.0116	2.0090	0.0070	2.0639	0.0241	1.9011	0.0176	1.9751	0.0082	1.9266	0.0084	1.9676
$\widehat{F}^{1}_{ycomp1}$	0.0050	0.9014	0.0015	0.9344	0.0022	0.9543	0.0025	0.9121	0.0032	0.8770	0.0031	0.7971	0.0024	0.8769	0.0015	0.8726
$\widehat{F}^{3}_{ycomp1}$	0.0022	0.4957	0.0036	0.4978	0.0034	0.4985	0.0020	0.4947	0.0031	0.4106	0.0015	0.3829	0.0028	0.3878	0.0016	0.4014
$\widehat{F}_{ycomp2}$	0.0048	1.0390	0.0040	1.0224	0.0034	1.0148	0.0041	1.0090	0.0050	1.0499	0.0068	1.0466	0.0029	1.0083	0.0017	1.0165
$\widehat{F}_{ycomp2red}$	0.0040	1.0260	0.0015	1.0113	0.0036	1.0013	0.0041	1.0074	0.0051	1.1080	0.0067	1.0399	0.0028	1.0070	0.0016	1.0163
$\widehat{F}_{ycomp3}$	0.0080	0.3432	0.0042	0.3652	0.0015	0.3614	0.0010	0.3578	0.0063	0.3297	0.0062	0.2690	0.0012	0.2510	0.0021	0.2755
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$PV = 80\%$								$PV = 90\%$							
	$n = 50$		$n = 75$		$n = 100$		$n = 125$		$n = 50$		$n = 75$		$n = 100$		$n = 125$	
$\widehat{F}_{HT}$	0.0051	1	0.0035	1	0.0028	1	0.0043	1	0.0040	1	0.0053	1	0.0020	1	0.0015	1
$\widehat{F}_{D}$	0.0047	0.9939	0.0034	0.9898	0.0030	0.9953	0.0043	1	0.0040	1	0.0053	1	0.0020	1	0.0015	1
$\widehat{F}_{R}$	0.0185	2.3237	0.0086	2.1132	0.0116	2.1872	0.0043	1	0.0212	2.1344	0.0049	2.2823	0.0066	2.0436	0.0015	1
$\widehat{F}_{CD}$	0.0075	0.9279	0.0069	0.9377	0.0088	0.9461	0.0080	0.9585	0.0118	0.9289	0.0086	0.9555	0.0071	0.9504	0.0058	0.9546
$\widehat{F}_{RKM}$	0.0046	1.0076	0.0036	0.9976	0.0028	1.0030	0.0049	1.0075	0.0052	1.0038	0.0051	1.0069	0.0019	0.9990	0.0015	1.0055
$\widehat{F}^{1}_{yc}$	0.0058	1.5264	0.0035	1.4115	0.0032	1.4692	0.0056	1.4925	0.0036	1.4239	0.0065	1.5371	0.0016	1.4149	0.0012	1.4524
$\widehat{F}^{3}_{yc}$	0.0058	1.4137	0.0034	1.3333	0.0028	1.3704	0.0040	1.3240	0.0045	1.3376	0.0051	1.3491	0.0014	1.2447	0.0016	1.3492
$\widehat{F}_{ycopt}$	0.0201	1.9932	0.0122	1.9845	0.0100	2.0354	0.0070	2.0639	0.0124	1.9534	0.0120	2.0248	0.0095	1.9861	0.0084	1.9676
$\widehat{F}_{ycoptred}$	0.0201	1.9932	0.0122	1.9845	0.0100	2.0354	0.0070	2.0639	0.0124	1.9534	0.0120	2.0248	0.0095	1.9861	0.0084	1.9676
$\widehat{F}^{1}_{ycomp1}$	0.0060	0.8920	0.0038	0.8553	0.0030	0.8820	0.0029	0.8816	0.0033	0.8552	0.0053	0.8771	0.0022	0.8628	0.0015	0.8726
$\widehat{F}^{3}_{ycomp1}$	0.0046	0.3992	0.0041	0.4064	0.0031	0.4007	0.0027	0.3878	0.0025	0.4018	0.0029	0.3974	0.0012	0.4116	0.0016	0.4014
$\widehat{F}_{ycomp2}$	0.0074	1.1463	0.0040	1.0465	0.0042	1.0581	0.0061	1.0329	1.6406	1.8742	0.0089	1.1484	0.0094	1.0596	0.0056	1.0771
$\widehat{F}_{ycomp2red}$	0.0242	3.7105	0.0041	2.3609	0.0055	2.6843	0.0061	1.0292	1.6809	1.7712	0.0221	1.2988	0.0107	1.8609	0.0059	1.2345
$\widehat{F}_{ycomp3}$	0.0089	0.5252	0.0064	0.2861	0.0020	0.2243	0.0023	0.1994	0.0040	1	0.0144	4.0845	0.0092	0.7507	0.0039	0.3692

An extensive simulation study was carried out, comparing the proposed estimators with other known ones for the distribution function. In this simulation study, the proposed estimators were found to be competitive in terms of bias reduction and mean square error compared with others such as the Rao–Kovar–Mantel estimator (which is known to present good properties). Furthermore, they present lower AVRG and AVRE values than the respective versions lacking calibration based on principal components. No single estimator performed best in all the situations considered, but $\widehat{F}_{ycomp3}$ was better than the rest in most of the scenarios considered. Moreover, this estimator can be used under any sample design and the weights thus obtained do not depend on the variable under study. For this reason, we recommend its use over the other estimators considered.

The proposed estimators can easily be adapted to address other population parameters of interest, such as quantiles and quantile-based measures of poverty, since under mild conditions, all of the estimators we discuss are genuine distribution functions.

Finally, like all research, this study is subject to some limitations. On the one hand, the methods proposed in [25] are not extended to consider the case of incomplete auxiliary information. Moreover, although we consider the method proposed by [25] to select the dimension of the principal components, thus ensuring that the estimators $\widehat{F}_{ycomp2}$; $\widehat{F}_{ycomp2red}$ and $\widehat{F}_{ycomp3}$ are monotone nondecreasing, we did not determine the existence or otherwise of conditions that guarantee monotony. Further work is needed to address this issue and, also, to analyse the performance of the proposed estimators when applied to the estimation of quantiles.

Data availability

The data that has been used is confidential.

Acknowledgements

Funding:
Grant A-FQM-170-UGR20 supported by Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía and FEDER.
Grant PID2019-106861RB-I00 supported by MCIN/ AEI /10.13039/501100011033
Grant CEX2020-001105-M supported by MCIN/AEI /10.13039/501100011033

Table 3
Average relative bias (AVRB) and average relative efficiency (AVRE) of the estimators compared for several sample sizes. Population: EUSILC.

	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$PV = 60\%$								$PV = 70\%$							
	$n = 150$		$n = 200$		$n = 250$		$n = 300$		$n = 150$		$n = 200$		$n = 250$		$n = 300$	
$\widehat{F}_{HT}$	0.0007	1	0.0016	1	0.0010	1	0.0023	1	0.0012	1	0.0016	1	0.0014	1	0.0024	1
$\widehat{F}_{D}$	0.0008	1.0007	0.0016	1.0024	0.0010	0.9987	0.0024	0.9981	0.0011	1.0006	0.0016	1.0008	0.0014	0.9998	0.0025	1.0014
$\widehat{F}_{R}$	0.0007	1.0162	0.0014	1.0011	0.0006	1.0305	0.0023	1.0295	0.0015	1.0165	0.0018	1.0122	0.0012	1.0197	0.0027	1.0107
$\widehat{F}_{CD}$	0.0238	1.273	0.0212	1.1570	0.0209	1.1732	0.0210	1.1828	0.0211	1.168	0.0194	1.0943	0.0206	1.1615	0.0187	1.1764
$\widehat{F}_{RKM}$	0.0021	1.0139	0.0017	1.0104	0.0009	1.0116	0.0022	1.0047	0.0016	1.0062	0.0014	1.0047	0.0014	1.0052	0.0024	1.0132
$\widehat{F}^{1}_{yc}$	0.0008	1.0818	0.0014	1.0473	0.0006	1.1183	0.0023	1.1134	0.0013	1.0916	0.0017	1.0790	0.0011	1.1012	0.0027	1.0847
$\widehat{F}^{3}_{yc}$	0.0007	1	0.0016	1	0.0010	1	0.0023	1	0.0012	1	0.0016	1	0.0014	1	0.0024	1
$\widehat{F}_{ycopt}$	0.0007	1.0152	0.0017	1.0124	0.0010	1.0052	0.0023	1.0040	0.0010	1.0049	0.0013	1.0092	0.0013	1.0009	0.0024	1.0094
$\widehat{F}_{ycoptred}$	0.0007	1.0152	0.0017	1.0124	0.0010	1.0052	0.0023	1.0040	0.0010	1.0049	0.0013	1.0092	0.0013	1.0009	0.0024	1.0094
$\widehat{F}^{1}_{ycomp1}$	0.0109	0.9896	0.0081	0.9575	0.0069	0.9788	0.0049	0.9190	0.0141	1.0124	0.0081	0.9910	0.0077	0.9558	0.0045	0.9055
$\widehat{F}^{3}_{ycomp1}$	0.0089	0.8343	0.0064	0.8133	0.0051	0.8233	0.0040	0.7652	0.0109	0.8412	0.0059	0.8235	0.0074	0.8298	0.0035	0.7427
$\widehat{F}_{ycomp2}$	0.0008	1.0818	0.0014	1.0473	0.0006	1.1183	0.0023	1.1135	0.0013	1.0916	0.0017	1.0790	0.0011	1.1012	0.0027	1.0847
$\widehat{F}_{ycomp2red}$	0.0008	1.0818	0.0014	1.0473	0.0006	1.1183	0.0023	1.1135	0.0013	1.0916	0.0017	1.0790	0.0011	1.1012	0.0027	1.0847
$\widehat{F}_{ycomp3}$	0.0003	0.8040	0.0025	0.8196	0.0014	0.7944	0.0034	0.8012	0.0026	0.7387	0.0003	0.7473	0.0032	0.7788	0.0012	0.7436
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$PV = 80\%$								$PV = 90\%$							
	$n = 150$		$n = 200$		$n = 250$		$n = 300$		$n = 150$		$n = 200$		$n = 250$		$n = 300$	
$\widehat{F}_{HT}$	0.0012	1	0.0005	1	0.0039	1	0.0013	1	0.0021	1	0.0004	1	0.0009	1	0.0023	1
$\widehat{F}_{D}$	0.0012	1.0003	0.0005	1.0002	0.0039	1.0004	0.0013	1.0013	0.0022	0.9996	0.0005	1.0002	0.0009	1.0005	0.0023	1.0002
$\widehat{F}_{R}$	0.0012	1.0172	0.0005	1.0170	0.0041	1.0174	0.0014	1.0092	0.0020	1.0223	0.0007	1.0169	0.0007	1.0130	0.0022	1.0169
$\widehat{F}_{CD}$	0.0199	1.1187	0.0201	1.1015	0.0190	1.1268	0.0182	1.1680	0.0217	1.1633	0.0200	1.1242	0.0208	1.1353	0.0212	1.2209
$\widehat{F}_{RKM}$	0.0026	1.0059	0.0007	1.0083	0.0038	1.0026	0.0014	1.0076	0.0006	1.0073	0.0005	1.0095	0.0009	0.9999	0.0022	1.0054
$\widehat{F}^{1}_{yc}$	0.0014	1.0861	0.0004	1.0835	0.0042	1.0880	0.0012	1.0612	0.0017	1.0992	0.0007	1.0932	0.0009	1.0774	0.0023	1.0884
$\widehat{F}^{3}_{yc}$	0.0012	1	0.0005	1	0.0039	1	0.0013	1	0.0021	1	0.0004	1	0.0009	1	0.0023	1
$\widehat{F}_{ycopt}$	0.0013	1.0158	0.0007	1.0126	0.0038	0.9995	0.0013	1.0057	0.0022	1.0074	0.0005	1.0130	0.0010	1.0001	0.0023	1.0039
$\widehat{F}_{ycoptred}$	0.0013	1.0158	0.0007	1.0126	0.0038	0.9995	0.0013	1.0057	0.0022	1.0074	0.0005	1.0130	0.0010	1.0001	0.0023	1.0039
$\widehat{F}^{1}_{ycomp1}$	0.02008	1.0727	0.0132	0.9647	0.0075	0.9612	0.0061	0.9691	0.0275	1.0461	0.0219	1.1334	0.0158	1.0625	0.0121	0.9798
$\widehat{F}^{3}_{ycomp1}$	0.0190	0.8851	0.0125	0.7991	0.0075	0.8560	0.0056	0.8027	0.0227	0.7805	0.0180	0.8547	0.0134	0.7950	0.0104	0.7560
$\widehat{F}_{ycomp2}$	0.0014	1.0861	0.0004	1.0835	0.0042	1.0880	0.0012	1.0612	0.0017	1.0992	0.0007	1.0932	0.0009	1.0774	0.0023	1.0884
$\widehat{F}_{ycomp2red}$	0.0014	1.0861	0.0004	1.0835	0.0042	1.0880	0.0012	1.0612	0.0017	1.0992	0.0007	1.0932	0.0009	1.0774	0.0023	1.0884
$\widehat{F}_{ycomp3}$	0.0031	0.9119	0.0019	0.8204	0.0005	0.7676	0.0012	0.7396	0.0145	1.5104	0.0145	1.0989	0.0063	1.0549	0.0040	1.3023

References

[1] 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.

[2] B. Bogin, T. Sullivan, Socioeconomic status sex, age, and ethnicity as determinants of body fat distribution for guatemalan children, Am. J. Phys. Anthropol. 69 (4) (1986) 527–535, http://dx.doi.org/10.1002/ajpa.1330690413.

[3] 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.

[4] M. Tellez-Plaza, A. Navas-Acien, C. Crainiceanu, E. Guallar, Cadmium exposure and hypertension in the 1999–2004 national health and nutrition examination survey (NHANES), Environ. Health Perspect. 116 (1) (2008) 51–56, http://dx.doi.org/10.1289/ehp.10764.

[5] R. Dickens, A. Manning, Has the national minimum wage reduced UK wage inequality? J. R. Stat. Soc. Ser. A 167 (4) (2004) 613–626, http://dx.doi.org/10.1111/j.1467-985X.2004.aeI2.x.

[6] S. Machin, A. Manning, L. Rahman, Where the minimum wage bites hard: Introduction of minimum wages to a low wage sector, J. Eur. Econ. Assoc. 1 (1) (2003) 154–180, http://dx.doi.org/10.1162/154247603322256792.

[7] 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.

[8] D. Morales, M. Rueda, D. Esteban, Model-assisted estimation of small area poverty measures: an application within the valencia region in Spain, Soc. Indic. Res. 138 (3) (2018) 873–900, http://dx.doi.org/10.1007/s11205-017-1678-1.

[9] 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.

[10] 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.

[11] 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.

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

[13] J. Mayor-Gallego, J. 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.

[14] 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.

[15] H. 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.

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

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

[18] 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.
[19] 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.
[20] 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.
[21] S. Martínez, M. Rueda, M. 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.
[22] K. McConville, F. Breidt, T. Lee, G.G. Moisen, Model-assisted survey regression estimation with the lasso, J. Surv. Stat. Methodol. 5 (2) (2017) 131–158, http://dx.doi.org/10.1093/jssam/smw041.
[23] J. Rao, A.C. Singh, Range restricted weight calibration for survey data using ridge regression, Pak. J. Stat. 25 (4) (2009).
[24] 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.
[25] H. Cardot, C. Goga, M.A. Shehzad, Calibration and partial calibration on principal components when the number of auxiliary variables is large, Statist. Sinica 24 (2017) 3–260, https://www.jstor.org/stable/44114370.
[26] 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.
[27] 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.
[28] D. Jackson, Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches, Ecology 74 (8) (1993) 2204–2214, http://dx.doi.org/10.2307/1939574.
[29] R. Johnson, D. Wichern, Applied Multivariate Statistical Analysis, sixth ed., Pearson Prentice Hall, New Jersey, 2007.
[30] P. Bardsley, R. Chambers, Multipurpose estimation from unbalanced samples, J. R. Stat. Soc. Ser. C 33 (3) (1984) 290–299, http://dx.doi.org/10.2307/2347706.
[31] 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.
[32] J. Rao, J. Kovar, H. Mantel, On estimating distribution functions and quantiles from survey data using auxiliary information, Biom. 77 (2) (1990) 365–375, http://dx.doi.org/10.2307/2336815.
[33] R. Chambers, R. Dunstan, Estimating distribution functions from survey data, Biom. 73 (3) (1986) 597–604, http://dx.doi.org/10.1093/biomet/73.3.597.
[34] F. Breidt, J. Opsomer, Local polynomial regression estimators in survey sampling, Ann. Statist. 28 (4) (2000) 1026–1053, https://www.jstor.org/stable/2673953.