SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Original Research Article

Statistical Methods in Medical Research
2023, Vol. 32(4) 760–772
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802231151211
journals.sagepub.com/home/smm
SAGE

A unified approach based on multidimensional scaling for calibration estimation in survey sampling with qualitative auxiliary information

J Fernando Vera[1], Carmen Cecilia Sánchez Zuleta[2], and Maria del Mar Rueda[1]

Abstract

Survey calibration is a widely used method to estimate the population mean or total score of a target variable, particularly in medical research. In this procedure, auxiliary information related to the variable of interest is used to recalibrate the estimation weights. However, when the auxiliary information includes qualitative variables, traditional calibration techniques may be not feasible or the optimisation procedure may fail. In this article, we propose the use of linear calibration in conjunction with a multidimensional scaling-based set of continuous, uncorrelated auxiliary variables along with a suitable metric in a distance-based regression framework. The calibration weights are estimated using a projection of the auxiliary information on a low-dimensional Euclidean space. The approach becomes one of the linear calibration with quantitative variables avoiding the usual computational problems in the presence of qualitative auxiliary information. The new variables preserve the underlying assumption in linear calibration of a linear relationship between the auxiliary and target variables, and therefore the optimal properties of the linear calibration method remain true. The behaviour of this approach is examined using a Monte Carlo procedure and its value is illustrated by analysing real data sets and by comparing its performance with that of traditional calibration procedures.

Keywords

Survey sampling, calibration weights, auxiliary information, categorical variables, multidimensional scaling, distance-based regression

1 Introduction

The use of sampling weights in survey data analysis is a widely used statistical tool in medical research.[1] In order to reduce the variance of the estimator of the total of a variable under a general sampling design in a finite population, the standard procedure is to use auxiliary information to adjust the sampling weights. This method was introduced by Deming and Stephan,[2] while the concept of *calibration or regression weights* was first mentioned by Huang and Fuller.[3] However, it was Deville and Särndal[4] who popularised and formalised the calibration method as an alternative to the well-known Horvitz-Thompson (*H-T*) estimator.[5] Although the calibration approach provides a general framework for constructing efficient estimators for other finite population parameters,[6–8] in this article, we focus on estimating finite population totals.[8]

Calibration weighting requires knowledge of the population total for a set of auxiliary variables, and the corresponding auxiliary information was given by the complete set of variables, if possible, or at least those in a sample.[9] The procedure is

[1]Department of Statistics and O.R., University of Granada, Granada, Spain
[2]Faculty of Basic Sciences, University of Medellin, Medellin, Colombia

Corresponding author:
J Fernando Vera, Department of Statistics and O.R., Faculty of Sciences, University of Granada, Avenida Fuentenueva S/N, 18071 Granada, Spain.
Email: jfvera@ugr.es

based on estimating the optimal weights that satisfy the calibration equation while remaining as close as possible to the basic H-T design weights (see Särndal,[10] and Beaumont and Rao for a recent overview[11]). Therefore, the calibration weights are estimated by minimising this difference (measured by a distance that need not be Euclidean), for which various optimisation procedures can be used.

Linear (based on Chi-squared distance) and *raking* (based on the Kullback-Leibler information) methods are commonly used in calibration optimisation.[4,12–14] If the auxiliary variables are quantitative, the linear method is most commonly used, since it is asymptotically equivalent to a GREG-type estimator, which is motivated by a linear regression model,[15] and if there is no strong correlation between the auxiliary variables, the solution can be obtained theoretically.[9]

In the linear method, when dummies are used in connection with qualitative variables, as usual, the set of auxiliary variables may expand considerably, even when only a few original categorical variables are involved. The size of this expansion would depend on the number of original categories in the auxiliary variables. This process usually generates a sparse, singular matrix, making it difficult to derive a solution using the linear method.[13]

Therefore, when the auxiliary information is related to qualitative variables or to a mixed set of variables, the raking calibration method is usually preferred, as it also provides a solution to the presence of negative weights which may arise with the linear method, whilst offering optimal values for weights that may be outliers.[9] This occurs because the exponential functional in its formulation becomes very steep, and so the new weights are likely to be very large, meaning that outliers may appear. The raking method of calibration on known marginal counts (incomplete post-stratification) gives estimators that can be described as generalised raking procedures (the classical raking ratio is a simple special case).

This approach is associated with the use of a post-stratification estimator,[16] and may present instability and/or an increased variance of the estimators. Since an increase in the number of levels of the qualitative auxiliary variables can lead to a greater number of homogeneous groups and, therefore, of post-strata, some strata in the sample may have few or no elements. In this case, the variance of the estimator increases and the estimator becomes unstable.[17,13]

As a running example, we consider study data concerning the *serum cholesterol* (Ch) levels recorded for patients *Heart Failure Prediction* dataset previously analysed by Detrano et al.[18] and later made freely available by Soriano,[19] in which 11 clinical characteristics were measured to predict heart disease events. A random sample of 92 patients provided data on categorical auxiliary variables regarding *gender*, *chest pain type*, *exercise-induced angina* and *heart disease*, with two, four, two and two categories, respectively. Also recorded were, *fasting blood sugar above 120 mg/dL*, *resting electrocardiogram results*, and *the slope of the peak exercise ST segment*, recoded as categorical variables with two, three and three levels, respectively. The present analysis is limited to estimating the population mean of the *serum cholesterol* variable.

Since the auxiliary information is qualitative, the linear method was applied by recoding the seven variables into 11 dummy variables: one each *gender*, *fasting blood sugar*, *exercise-induced angina* and *heart disease*, two for *resting electrocardiogram results* and *slope of the peak exercise ST segment*, and three for *chest pain type*. For this data set, singularity problems were encountered with the linear method, for which no solution was obtained. In an alternative approach, the raking method was applied using a multiway contingency table of $2^4 \times 3^2 \times 4 = 576$ cells or post-strata. However, many of these 576 post-strata had only a few units or were empty, which led to divergence (see e.g. Guggemos and Tillé[20]).

Nascimento Silva and Skinner[21] experimentally showed that adding auxiliary variables reduces the error of calibrated estimates, up to a point, after which the number of auxiliary variables becomes very large and the mean squared error (MSE) increases. If too many auxiliary variables are used, the bias of the calibrated estimator increases and can become nonnegligible compared to the variance (over-calibration). Chauvet and Goga[22] have analysed the asymptotic efficiency of the calibration estimator with high-dimensional auxiliary data sets. They have shown that this may suffer from an additional variability that may not be neglected in certain conditions, given a theoretical justification of the increase of the variance of the calibration estimator. Additionally, the presence of missing data in the auxiliary information is also a major problem to consider in this framework.[23]

Various procedures have been suggested for variable selection when the auxiliary information is quantitative. One of the first was proposed by Nascimento Silva and Skinner,[21] who computed the MSE for all possible subsets of quantitative auxiliary variables and then chose the one producing the smallest MSE. Later, Clark and Chambers[24] used forward, backward and stepwise selection based on the difference between the MSE of the prediction for two nested sets of variables. Alternatively, the least absolute shrinkage and selection operator[25] might be considered for selecting the best subsets. Once the best set of regressors has been selected, the calibration is performed on these variables alone. Another approach to consider is that of penalised calibration,[20,26,27] which takes account of auxiliary information by attaching more or less importance according to its presumed explanatory power for the variable of interest. In a different way, Cardot et al.[28] suggested applying principal component analysis for quantitative auxiliary variables in order to achieve a strong dimension reduction and to perform the calibration only on the first principal component. Chauvet and Goga[22] have proposed a bootstrap criterion for the choice of calibration variables. Their simulation study suggests that the proposed method leads to a more parsimonious number of calibration variables, with associated weights of smaller variation without variance inflation.

In this article, we propose a procedure based on the use of multidimensional scaling (MDS) together with an appropriate dissimilarity measure to address the auxiliary information obtained from categorical variables.[29] Our proposal is based on projecting the target variable into the space spanned by the columns obtained by classical MDS, which play the role of independent predictive continuous auxiliary variables, combining this approach with the linear calibration method. The new set of auxiliary variables is related to the original set of qualitative or mixed variables through the Euclidean distances, while, as shown in Section 3.1, maintaining the implicit underlying assumption of linearity in the linear calibration procedure and, therefore, the optimal properties of this estimator. The auxiliary variables defined by the proposed procedure are related to a set of positive eigenvalues, which makes it possible to obtain a feasible estimator of the calibration weights as it is not affected by the singularity problem of the linear calibration procedure.

The next section describes the calibration procedures most commonly employed to estimate the total of a variable in a population when the auxiliary information is qualitative or mixed. Section 3 then describes the MDS procedure to obtain the new auxiliary variables for linear calibration, along with an appropriate dissimilarity measure. The linear relation of these new variables with the target variable is shown on the basis of an existing relation with the original variables, after recoding, and a procedure to determine the number of new auxiliary variables to be considered is then described. In Section 4, we analyse the behaviour of the model using a Monte Carlo experiment and the performance, as reflected in empirical data, is compared with that obtained by traditional calibration procedures. In the final section, we present and discuss the results obtained.

2 Calibration with auxiliary qualitative information

Let Y be a target variable and s a sample of size n from a finite population $\mathcal{U}$ of size N, under a sampling design in which $\pi_k = P[k \in s] > 0$ and $\pi_{k,l} = P[k \in s, l \in s] > 0$ denote the first and second-order probabilities respectively for $k, l \in s$. Let $\mathbf{X}_1^*, \mathbf{X}_2^*, \ldots, \mathbf{X}_p^*$ be a set of p auxiliary variables related to the target variable Y, whose values are known for the entire population, or alternatively only in the sample if the population totals for these variables are also known.

The calibration procedure is employed to find an efficient estimator for the total of Y denoted by $t_y = \sum_{k \in \mathcal{U}} y_k$, in which y_k represents the values of Y observed in the unit $k \in \mathcal{U}$, which are assumed to be known for all $k \in s$. A classical unbiased estimator of the total t_y is the H-T estimator, which uses the information in the sample units y_k, weighted by the inverse of the probabilities of inclusion $d_k = 1/\pi_k$, to calculate the estimator,

$$\hat{t}_{y_{HT}} = \sum_{k \in s} d_k y_k. \tag{1}$$

The calibration method outperforms *H-T* by using auxiliary information related to the target variable, of which the population total is known. Our aim in applying this method is to find new weights w_k that are as close as possible to the d_k, $k \in s$ (to preserve the property of unbiasedness), while satisfying the calibration equation constraint

$$\sum_{k \in s} w_k \mathbf{x}_k^* = \mathbf{t}_\mathbf{x}^* = (t_{\mathbf{x}_1^*}, t_{\mathbf{x}_2^*}, \ldots, t_{\mathbf{x}_p^*})^T, \tag{2}$$

where $\mathbf{x}_k^* = (x_{k1}^*, x_{k2}^*, \ldots, x_{kj}^*, \ldots, x_{kp}^*)^T$ is the observed vector of auxiliary information in the unit k, and $\mathbf{t}_\mathbf{x}^* = \sum_{k \in \mathcal{U}} \mathbf{x}_k^*$ is the vector of the totals of the auxiliary variables within the population.

The solution to equation (2) is not unique. For instance, given $\mathbf{x}_k^* = (1, x_k^*)^T$, in which x_k^* is a binary categorical variable,[30,12] the weights are adjusted to the population size and the number of elements in each category. In this situation, many solutions will satisfy (2) for $n > p$. Therefore, we must identify the w_k values that best approximate the *H-T*'s weights. In this respect, Deville and Särndal[4] introduced a distance function $G(., d_k)$ that under certain regularity conditions (see also Devaud and Tillé,[9] Wu and Thompson,[15] for further details) solves the following optimisation problem

$$w_k(s) = \min_{w_k \in \mathcal{A}} \sum_{k \in s} G(w_k, d_k), \tag{3}$$

where $\mathcal{A} = \{\mathbf{v} \in \mathbb{R}^n | \mathbf{X}^{*T}\mathbf{v} = \mathbf{t}_\mathbf{x}^*\}$ is the set of all the possible weights that satisfy the condition of perfect estimation given in (2). Since $G(\cdot, d_k)$ is a continuously differentiable function, and denoting by $\boldsymbol{\lambda}$ the vector of Lagrange's multipliers related to the constraints given in (2), the minimisation problem is reduced to that of solving the equation system

$$g(w_k, d_k) - \boldsymbol{\lambda}^T \mathbf{x}_k^* = 0 \quad \forall k \in s, \tag{4}$$

where $g(w_k, d_k)$ is the derivative of $G(\cdot, d_k)$ with respect to w_k, $k \in s$. Furthermore, if $\boldsymbol{\lambda}^T \mathbf{x}^*_k$ is a feasible point in the image of $g(\cdot, d_k)$, the solution to equation (4) is unique,[9] and is given by

$$w_k = d_k F_k(q_k \boldsymbol{\lambda}^T \mathbf{x}^*_k), \tag{5}$$

where $d_k F_k(\cdot)$ is the inverse function of $g(\cdot, d_k)$ (see e.g. Deville and Särndal[4]), which is an increasing function, where $F_k(0) = 1$ and $F'_k(0) = q_k$, and where the q_k are positive weights not related to d_k. Although unequal weights are also used to assign different levels of importance to the units, in most cases, $q_k = 1$, $k \in s$, and equation (5) adopts the expression

$$w_k = d_k F_k(\boldsymbol{\lambda}^T \mathbf{x}^*_k). \tag{6}$$

The $\boldsymbol{\lambda}$ values are found by substituting the weights w_k (6) into the calibration equation (2), and the calibration estimator is then expressed by

$$\hat{t}_{yw} = \sum_{k \in s} w_k y_k = \sum_{k \in s} d_k F_k(\boldsymbol{\lambda}^T \mathbf{x}^*_k) y_k. \tag{7}$$

2.1 Linear and raking calibration methods

Deville and Särndal[4] described various distance functions that can be used to carry out the necessary optimisation. Two of the functions most commonly used are the chi-square distance, which is related to the linear method, and the pseudo-entropy distance, related to the raking method.

The linear method is most commonly used, despite the risk of its producing negative weights. In this case, both the g and the F functions are linear and the chi-square distance function is given by

$$G_{linear}(w_k, d_k) = \frac{(w_k - d_k)^2}{2 q_k d_k}. \tag{8}$$

Here $F(u) = 1 + q_k u$ and $q_k = 1$, which gives rise to the following expression for the weights in the linear case,

$$w_k = d_k - d_k \mathbf{x}^{*T}_k \left(\sum_{\ell \in s} d_\ell \mathbf{x}^*_\ell \mathbf{x}^{*T}_\ell \right)^{-1} \left(\hat{t}_{\mathbf{x}^* HT} - \mathbf{t}_{\mathbf{x}^*} \right), \tag{9}$$

where $\hat{t}_{\mathbf{x}^* HT} = \sum_{k \in s} \mathbf{x}^*_k d_k$ is the *H-T* estimator for the total of the auxiliary variables. The expression for the weights found in (9) together with the expression of the calibration estimator (7) leads to the linear regression estimator given by

$$\hat{t}_{yw} = \hat{t}_{y_{HT}} + \left(\mathbf{t}_{\mathbf{x}^*} - \hat{t}_{\mathbf{x}^* HT} \right)^T \left(\sum_s d_k \mathbf{x}^*_k \mathbf{x}^{*T}_k \right)^{-1} \sum_s d_k y_k \mathbf{x}^*_k. \tag{10}$$

The raking method is especially useful when calibration is performed over a known marginal count in a frequency table, or for qualitative auxiliary information.[12] Although it circumvents the problem of negative weights (only positive weights are obtained), outlier values may appear.[9] The distance function for the raking method is given by

$$G_{raking}(w_k, d_k) = \frac{1}{q_k} \left[w_k \log \frac{w_k}{d_k} - w_k + d_k \right], \tag{11}$$

where $F(u) = \exp(q_k u)$, $u = \boldsymbol{\lambda}^T \mathbf{x}^*_k$ and $q_k = 1$; then the weights are given by $w_k = d_k \exp(q_k \boldsymbol{\lambda}^T \mathbf{x}^*_k)$. Again, λ_k values are found using this expression in (2), and the following equation is solved iteratively,

$$\sum_{k \in s} d_k \mathbf{x}^*_k \exp(q_k \boldsymbol{\lambda}^T \mathbf{x}^*_k) = \mathbf{t}_{\mathbf{x}^*}. \tag{12}$$

2.2 Qualitative auxiliary information

In most experimental situations, it is quite common for a significant amount of auxiliary information to be obtained from qualitative variables, which are usually recoded using dummy variables.[31] This auxiliary information is generally handled in terms of counts in a multi-way contingency table. In a mixed situation in which the auxiliary information is given by both quantitative and qualitative variables, the quantitative ones are usually categorised, and the frequency table is formed using the qualitative and the newly categorised variables.

Any qualitative variable **A** with I categories induces a partition on the population. As is well known, each category A_i of **A** can be recoded by the values of a set of I dummy variables, taking the value of one in the ith position and zero otherwise. Hence, for the k-th unit, the auxiliary information $\mathbf{a}_k$ can be defined by a vector of dimension I in which all entries are zero-valued, except for the category to which it belongs, which has the value one. Each level of the variable in the sample can be related to a post-stratum, and the estimator for the total is given in terms of the sum of the estimators in each post-stratum.

When several qualitative variables are present, we may consider a post-stratification process related to the corresponding multi-way contingency table. Let p be the number of auxiliary variables and I_r the number of levels in the r-th variable, $r = 1, \ldots, p$, and consider the cross-classified table of dimension $I_1 \times I_2, \times \cdots \times, I_p$. The entries in the table are the number of study units in the sample that satisfy the characteristics associated with each cell, which induces a partition in the population. If the population totals in each cell are also known, the raking calibration method can be used to estimate the weights. However, the estimation procedure is usually unstable when there are cells with small or null frequencies, a situation that more commonly occurs with higher values of p. Although this problem can sometimes be addressed by collapsing the cells,[30] this procedure does not ensure an acceptable solution or even mean that the problem cannot be solved. Although combining the cells might offer a solution to the problem, different criteria for grouping can give rise to different solutions, since the specificity of the categories involved decays in favour of the joint information of the resulting group. In addition, the resulting number of groups must be selected, which also represents a major problem. In short, determining the most appropriate solution depends to a large degree on the skill and experience of the researcher.

When the total population of each cell is not known, but the total of the original qualitative variables is known, together with the auxiliary values in a sample, the calibration method can be applied to the marginal totals of the table. This procedure can also be used when the previous procedure does not obtain a solution, but at the expense of an increase in the variability of the estimator. In this situation, the convergence of the raking calibration method may also fail, since the problem of cells with small frequencies is even more evident within the sample data, and the problem of small frequencies also occurs in the marginal ones. Therefore, it would be desirable to have an estimation procedure that is at least stable in these situations in order to perform the calibration with qualitative variables.

3 MDS-based linear calibration (MDSC)

The linear calibration method assumes a linear relationship between the continuous target variable and the auxiliary variables. If the auxiliary information is given by qualitative, binary and/or continuous variables, and we wish to relate them to a continuous variable $\boldsymbol{Y}$, the classical linear model would not be appropriate. Accordingly, various alternatives have been proposed. The usual procedure, however, is to subject all the qualitative variables to a given scoring system and then consider them as continuous. Another approach would be to follow the rationale of the distance-based regression model proposed by Cuadras and Arenas.[32] In this case, the all-purpose measure of similarity given by Gower[33] is the most appropriate,[34]

$$s_{ij} = \frac{\sum_{\ell=1}^{b_1}\left(1 - \frac{|x^*_{i\ell} - x^*_{j\ell}|}{R_\ell}\right) + a + \alpha}{b_1 + (b_2 - d) + b_3}, \tag{13}$$

where b_1 is the number of continuous variables, R_l is the range of the l-th continuous variable, a and d are the number of positive and negative matches, respectively, for the b_2 dichotomous variables, and α is the number of matches for the b_3 categorical non-binary variables. Then, $d^2_{ij} = 1 - s_{ij}$ represents the squared Euclidean distance between the units and therefore $\boldsymbol{D} = (d_{ij})$ is the matrix of Euclidean distances between the units related to an unknown configuration.[33]

This coefficient is equivalent to the simple matching coefficient when all variables are qualitative, and is directly proportional to the squared Euclidean distance between the sample units in dummy variable forms, that is, after the qualitative variables are coded into dichotomous ones. In consequence, the coefficient is equivalent to the squared Euclidean distance between vectors of length equal to the sum of the categories for the p categorical variables. For mixed variables, the coefficient can be calculated (after standardisation) using the Euclidean distance between vectors of length equal to the number of continuous variables plus the number of dichotomous variables resulting from the above described recoding for polychotomous categorical variables, plus the number of dichotomous variables, if any (which are each coded as a single dummy variable). Henceforth, we use $\boldsymbol{X}^*$ to represent the extended $n \times p$ centred matrix of the auxiliary information from which Gower's generalised coefficient is calculated, in terms of the Euclidean distances, in which $p < <(n-1)$ is the total number of auxiliary variables after recoding. In the following, we examine how the linear relation between the auxiliary and the target variables is preserved.

3.1 The linear relation in centred orthogonal form using MDS

For $\boldsymbol{X}^*$ representing the $n \times p$ matrix with the observed auxiliary information for a set of p variables $\boldsymbol{X}_1^*, \cdots, \boldsymbol{X}_p^*$, is denoted by $\boldsymbol{D}(\boldsymbol{X}^*)$ the $n \times n$ matrix of the Euclidean distances between the n units in the sample. It is assumed without loss of generality that $\boldsymbol{X}^*$ is centred, that is, $\boldsymbol{X}^* = \boldsymbol{H}\boldsymbol{X}^*$, where $\boldsymbol{H} = \boldsymbol{I} - \mathbf{1}\mathbf{1}^T/n$, $\boldsymbol{I}$ is the identity matrix and $\mathbf{1}$ is a $n \times 1$ vector of ones.

If $\boldsymbol{Y}$ is linearly related to $\boldsymbol{X}^*$, as would be desirable for the linear calibration method, we can assume the following expression,

$$\boldsymbol{Y} = \beta_0\mathbf{1} + \boldsymbol{X}^*\boldsymbol{\beta}_1 + \boldsymbol{e}, \tag{14}$$

where β_0 is an unknown scalar and $\boldsymbol{\beta}_1$ is an unknown parameter vector, and we assume that $E(\boldsymbol{e}) = 0$, and $E(\boldsymbol{e}\boldsymbol{e}^T) = \sigma^2\boldsymbol{I}$, with σ^2 unknown. Since $\boldsymbol{X}^*$ is centred, it follows that

$$n\boldsymbol{S}_{X^*} = \boldsymbol{X}^{*T}\boldsymbol{X}^* = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T, \tag{15}$$

where $\boldsymbol{S}_{X^*}$ is the covariance matrix of $\boldsymbol{X}^*$, $\boldsymbol{V}$ represents the eigenvectors in columns and $\boldsymbol{\Lambda}$ is the diagonal matrix of the non-zero eigenvalues, $\lambda_1 \geq \cdots \geq \lambda_m > 0$, with $m \leq p$, and $m = rank(\boldsymbol{X}^*)$. Also, from the Euclidean distance matrix $\boldsymbol{D}(\boldsymbol{X}^*)$ we see that $\boldsymbol{B} = -0.5\boldsymbol{H}\boldsymbol{D}^2\boldsymbol{H}$ is the corresponding matrix of scalar products for the centred configuration $\boldsymbol{X}^*$, with $rank(\boldsymbol{B}) = m$.[35] Then, $\boldsymbol{B}$ is the symmetric and positive semi-definite (p.s.d.) and, therefore, $\boldsymbol{B} = \boldsymbol{X}^*\boldsymbol{X}^{*T} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \boldsymbol{X}\boldsymbol{X}^T$, where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$, and where $\boldsymbol{U}$ is the $n \times m$ matrix of eigenvectors.

Hence, we consider the singular value decomposition of $\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$, from where $\boldsymbol{X}^*\boldsymbol{V} = \boldsymbol{X}\boldsymbol{\Lambda}^{1/2}$, and therefore $\boldsymbol{Y}$ is linearly related to $\boldsymbol{X}$ through the expression

$$\boldsymbol{Y} = \beta_0\mathbf{1} + \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{e}, \tag{16}$$

where $\boldsymbol{\gamma} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{V}^T\boldsymbol{\beta}_1$. Therefore, we need only know $\boldsymbol{D}(\boldsymbol{X}^*)$, in order to obtain $\boldsymbol{X}$ using MDS (regardless of $\boldsymbol{X}^*$, or even if some data are missing), in such a way that the underlying linear relationship is preserved. Furthermore, $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{\Lambda}$, which means the linear relation is expressed in centred orthogonal form.

As noted above, the original linear relation between the auxiliary information and the target variable, if any, is preserved using Gower's general dissimilarity coefficient, which is equivalent to a squared Euclidean distance, and the auxiliary information can be expressed through a new set of the same number of $m < p$ auxiliary variables from the MDS-based procedure, which are also uncorrelated. Furthermore, as with the distance-based regression model, if 0, $\lambda_1 \geq \cdots \geq \lambda_r$ represents a suitable set of eigenvalues, the latter related to the r columns of $\boldsymbol{X}$, denoted by $\boldsymbol{X}_{(r)} = (X_1, \cdots, X_r)$, with $r \leq m$, then the following linear model can be considered for the purposes of calibration:

$$\boldsymbol{Y} = \beta_0\mathbf{1} + \boldsymbol{X}_{(r)}\boldsymbol{\gamma}_{(r)} + \boldsymbol{e}_{(r)}, \tag{17}$$

where $\boldsymbol{\gamma}_{(r)}$ and $\boldsymbol{e}_{(r)}$ are the corresponding conformable matrices.

Even with a large sample size n, the number of MDS-based auxiliary variables does not exceed the total number of original quantitative variables plus the number of dummies, after recoding the qualitative variables. Although the problem of selecting the dimensionality with which to decide the optimal number of auxiliary variables and thus to estimate the weights persists, the proposed procedure also offers a natural solution based on MDS.

3.2 Calibration using the MDS-based configuration

The results shown in the previous section exactly define the linear relationship between the response variable $\boldsymbol{Y}$ and the set of auxiliary quantitative variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r$ obtained by the proposed procedure based on MDS, after performing an appropriate recoding of the auxiliary information. Therefore, the linear calibration method can be performed on this set of variables to estimate the new weights, while the asymptotic properties of the estimators are preserved.

For this new set of r auxiliary variables given from MDS, the calibration weights (9) can be written as

$$w_k^{MDS}(r) = d_k - d_k\mathbf{x}_{(r)_k}^T\left(\sum_{\ell\in s} d_\ell\mathbf{x}_{(r)\ell}\mathbf{x}_{(r)_\ell}^T\right)^{-1}\left(\hat{\mathbf{t}}_{\mathbf{x}_{(r)}HT} - \mathbf{t}_{\mathbf{x}_{(r)}}\right), \text{with k} \in \text{s}, \tag{18}$$

where $\hat{\mathbf{t}}_{\mathbf{x}_{(r)HT}} = \sum_{k\in s} d_k \mathbf{x}_{(r)_k}$ is the estimator of *H-T* of the total $\mathbf{t}_{\mathbf{x}_{(r)}}$ for the auxiliary variables, and $\mathbf{t}_{\mathbf{x}_{(r)}} = (0, 0, \ldots, 0)$, since the MDS configuration is centred.

Finally, the calibration estimator (GREG) for the total t_y is given by

$$\begin{aligned}\hat{t}_{yw}^{MDS}(r) &= \sum_{k\in s} w_k^{MDS}(r) y_k \\ &= \hat{t}_{yHT} - \left(\hat{\mathbf{t}}_{\mathbf{x}_{(r)}HT} - \mathbf{t}_{\mathbf{x}_{(r)}}\right)^T \left(\sum_{k\in s} d_k \mathbf{x}_{(r)_k} \mathbf{x}_{(r)_k}^T\right)^{-1} \sum_{k\in s} d_k \mathbf{x}_{(r)_k} y_k.\end{aligned} \tag{19}$$

Under the asymptotic framework of Isaki and Fuller,[36] and the usual assumptions to establish the consistency of the GREG estimator,[4,28] the estimator $\hat{t}_{yw}^{MDS}(r)$ given in (19) verifies that:

1. $(\hat{t}_{yw}^{MDS}(r) - t_y)/N = o_p(n^{-1})$. In other words, the estimator $\hat{t}_{yw}^{MDS}(r)$ is design-consistent;
2. The asymptotic variance of the estimator is given by

$$AV(\hat{t}_{yw}^{MDS}(r)) = \sum_{k,l\in U} (\pi_{kl} - \pi_k \pi_l) d_k e_{(r)k} d_l e_{(r)l},$$

where $e_{(r)k}$ denotes the residual of unit k given by model (17).

These results are derived from the properties of the linear calibration estimator, together with the fact that the proposed estimator preserves the linear relation. In addition, the above formula allows us to describe estimators for the variance, which is important to quantify the error and to determine whether the proposed method actually reduces the variance. Since $e_{(r)k}$ is based on population coefficients, it cannot be computed and must be replaced by the sample-based residual $\tilde{e}_{(r)k} = y_k - \beta_{s0} + \mathbf{x}_{(r)k}\gamma_{(sr)}$ where β_{s0} and $\boldsymbol{\gamma}_{(sr)}$ are the weighted sample coefficients. Thus, two variance estimators of $\hat{t}_{yw}^{MDS}(r)$ are given by

$$\hat{V}_1(\hat{t}_{yw}^{MDS}(r)) = \sum_{k,l\in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} d_k \tilde{e}_{(r)k} d_l \tilde{e}_{(r)l}, \tag{20}$$

and

$$\hat{V}_2(\hat{t}_{yw}^{MDS}(r)) = \sum_{k,l\in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} w_k^{MDS}(r) \tilde{e}_{(r)k} w_l^{MDS}(r) \tilde{e}_{(r)l}. \tag{21}$$

Diverse criteria can be employed to determine the optimal number of variables in the MDS-based calibration procedure. It would be desirable to select a number of auxiliary variables explaining a suitable proportion of variation in the data set, while ensuring a high proportion of the calibration weights remain positive. Since the MDS solution is associated with the eigenvectors related to the largest (positive) eigenvalues, the problem is reduced to that of selecting the first r dimensions in the MDS-based configuration, for an appropriate value of r. It is important to highlight that (17) does not suffer from a singularity problem, since the r column vectors in the MDS configuration are orthogonal, and therefore the inverse matrix always exists. Thus, the proposed procedure ensures a feasible solution to the problem of estimating the calibration weights.

3.3 Overview and implementation

The algorithm has been implemented in *R* to show the performance of the proposed MDS calibration estimator. An overview of this algorithm is provided in Figure 1[1]. When data for the entire population are available, five thousand random samples are selected ($Nsamples = 5000$) and the relative mean square error with respect to the H-T estimator is also calculated.

4 Experimental results

In this section, we present an application of the MDS-based calibration procedure to the heart failure prediction dataset and analyse its properties using a Monte Carlo experiment; in each case, we compare the results obtained with those found with the classical linear and raking calibration.

```
MDS-based linear calibration algorithm.
Input: Data.
Initialise: Obtain the centered extended data matrix $\mathbf{X}^*$.
Calculate: $\boldsymbol{D}(\boldsymbol{X}^*)$ the $n \times n$ matrix of dissimilarities using (13).
Estimate: $\mathbf{X}$ and $m$ using classical MDS in full dimension,
Set: $\mathbf{t}^*_{\mathbf{x}}$ the vector of totals, $\lambda_i$ the positive eigenvalues.
    $Nsamples = 5000$. If data is observed in a sample, $Nsamples = 1$
    $g = 0$, $i = 1$
While $i \leq Nsamples$
    do $g = g + 1$ Count the samples generated
     Select $MuestMDS_i$, and $Y_i$ samples
     Calculate: $\hat{t}_{y_{HT}}$ (1)
    For: each dimension $r$ until the number of positive eigenvalues $m$, calculate:
        $w_k^{MDS}(r)$ with (18)
        $\hat{t}_{yw}^{MDS}(r)$ using (19)
        $R(\hat{t}_{yw}^{MDS}(r))$ the relative estimation error of $\hat{t}_{yw}^{MDS}(r)$ using (22)
        the proportion of positive weights $w_k^{MDS}(r)$.
     End
     do $i = i + 1$
End
Calculate: $MSEr_{MDS}$ relative mean square error of $\hat{t}_{yw}^{MDS}(r)$ using (23).
Print $\hat{t}_{y_{HT}}$, $\hat{t}_{yw}^{MDS}(r)$, $MSEr_{MDS}$.
        proportion of positive weights for each dimension.
```

Figure 1. Multidimensional scaling (MDS)-based linear calibration algorithm.

4.1 Heart failure prediction dataset

Here, we analyse the data introduced in Section 1. The proposed model was applied to the survey sample of $n = 92$ respondents from a population total of $N = 918$ patients. The auxiliary variables considered are categorised as: F=female or M=male for *gender*; ASY = Asymptomatic, ATA = Atherosclerosis, NAP = Non-Anginal Pain or TA = Typical Angina for *chest pain type*; T = true or F = false for *fasting blood sugar above 120 mg/dl*; N=normal, LVH = showing probable or definite left ventricular hypertrophy by Estes' criteria, or ST = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression above 0.05 mV) for *resting electrocardiogram results*; N = no or Y = yes for *exercise-induced angina*; down, flat or up-sloping for *the slope of the peak exercise ST segment*; 0 = below 50% or 1 = above 50% diameter narrowing for *heart disease*.

4.1.1 Traditional calibration performance

The linear calibration procedure was first performed after recoding the qualitative variables, as described in Section 1. However, the matrix $\sum_{\ell \in s} d_\ell \mathbf{x}^*_\ell \mathbf{x}^{*T}_\ell$ in (9) and (18) is singular in this data set, and no solution was found for this procedure.

The raking calibration method was applied to this data set after cross-tabulation of the seven original variables. In this case, many empty cells were found in the table, in particular those cells that have a cross-relationship between any level of the variable *type of chest pain*, and the female level of the variable *gender*. In fact, many of the marginal frequencies of these levels were equal to zero. Moreover, collapsing the cells did not solve the problem, since the number of marginal cells with low frequencies remained high, and the calibration estimator presented divergence.[20]

4.1.2 Calibration using the MDS-based procedure

The MDS-based calibration procedure was then performed using Gower's coefficient. In this approach, the population mean estimator of serum cholesterol (Ch) was calculated with linear calibration using from one to 11 MDS-based auxiliary variables ($r = 1, \ldots, 11$ dimensions). Since the MDS solution is based on the Euclidean distances, all the eigenvalues are positive. For all MDS solutions up to 11 dimensions, the estimated population mean values of the variable (Ch) are presented in Table 1, together with their 95% confidence intervals, the coefficient of variation ($C_v = sd_w/\overline{w}$) of the weights,

Table 1. Results obtained with the multidimensional scaling (MDS)-based calibration procedure for the variable *Cholesterol* up to dimension $r = 11$, showing the mean estimated $\hat{\mu}_{yw}^{MDS}$, the coefficient of variation of the weights (CV), the goodness of fit (GOF) in MDS, and the relative estimation error of the MDS estimator with respect to the H-T estimator ($R(\hat{\mu}_{yw}^{MDS}(r))$). The two rows corresponding to the lowest relative estimation errors are indicated in bold.

r	$\hat{\mu}_{yw}^{MDS}$	Linf	Lsup	CV	GOF	$R(\hat{\mu}_{yw}^{MDS}(r))$
1	210.62	210.44	210.80	0.1703	0.36	1.44
2	207.46	207.28	207.64	0.1947	0.47	0.77
3	208.93	208.75	209.11	0.1964	0.57	1.06
4	209.01	208.83	209.18	0.1971	0.67	1.08
5	208.99	208.81	209.17	0.1972	0.75	1.07
6	209.13	208.95	209.30	0.1958	0.82	1.10
7	***207.53***	***207.36***	***207.70***	***0.2143***	***0.87***	***0.79***
8	207.76	207.59	207.93	0.2188	0.92	0.83
9	208.18	208.01	208.35	0.2840	0.96	0.91
10	***206.23***	***206.06***	***206.40***	***0.2996***	***0.98***	***0.57***
11	223.29	223.03	223.54	74.2673	1.0	6.19

the explained variability (goodness of fit (GOF)) in classical MDS, and the relative estimation error of the MDS estimator, $R(\hat{\mu}_{yw}^{MDS}(r))$, with respect to the H-T estimator

$$R(\hat{\mu}_{yw}^{MDS}(r)) = \frac{(\hat{\mu}_{yw}^{MDS}(r) - \mu_y)^2}{(\hat{\mu}_{y_{HT}} - \mu_y)^2}. \tag{22}$$

With the linear calibration procedure, all the estimated weights for the MDS-based auxiliary variables were positive in all dimensions except in 11 dimensions, in which only 69.5% of positive weights were obtained. Regarding the dispersion of the weights, in each dimension up to $r = 6$ the coefficient of variation was below 0.2, for $r = 7$, 8 it was below 0.22, for $r = 9$, 10 it was below 0.3, and only in the last dimension was it above one.

The lowest relative estimation error of the MDS estimator was observed for $r = 10$, followed by $r = 7$, which corresponds to an explained proportion of the variability of 98.36% and 87.2%, respectively. Since all the weights were positive in both dimensions, we selected ten auxiliary variables for the linear calibration procedure. Thus, unlike both traditional calibration methods, and despite the absence of a strong linear relationship between the auxiliary variables and the target variable, the proposed MDS-based procedure produced a good estimator of the mean of the variable Cholesterol, according to which the relative estimation error of the MDS estimator was equal to 0.57. Note that in this case, the estimation error was 0.169 with a confidence level of 95%, a value that is less than the mean estimation error (0.183) and less than the median of the errors (0.177).

4.2 Monte Carlo results

A Monte Carlo experiment was conducted to evaluate the performance of the MDS-based calibration procedure for the *serum cholesterol* and the *maximum heart rate achieved* variables, and to compare the results obtained with those of the traditional linear and raking calibration procedures. In this experiment, 5000 samples of size $n = 92$ were drawn by a simple random sampling design without replacement. For each sample, the estimator of the population mean was calculated. Then, the goodness of fit of the estimated values was determined by examining the relative mean square error (MSEr) in relation to the H-T estimator,

$$MSEr = R(\hat{\mu}_{cal}) = \frac{\sum_{b=1}^{B} \left(\hat{\mu}_{cal}^{(b)} - \mu_y\right)^2}{\sum_{b=1}^{B} \left(\hat{\mu}_{y_{HT}}^{(b)} - \mu_y\right)^2}, \tag{23}$$

where μ_y denotes the population mean of the target variable, and for each b-th Monte Carlo sample, $\hat{\mu}_{cal}^{(b)}$ is the value of the calibration estimator both for the classical ($\hat{\mu}_{yw}$) and the MDS-based ($\hat{\mu}_{yw}^{MDS}$) procedures, and $\hat{\mu}_{y_{HT}}^{(b)}$ is the estimated mean value of the $H - T$ estimator. The MSEr values obtained by the traditional linear calibration method were 0.9605031 and 0.8681743 for the *Ch* and *MaxHR* variables, respectively, while those by the raking calibration procedure were of 0.9970190 and 0.8650021, respectively. Table 2 shows the results obtained in each dimension by the MDS-based auxiliary variables using both calibration procedures.

Table 2. Relative mean square error results for the linear and raking procedures, both for the cholesterol (Ch) and the maximum heart rate achieved (MaxHR) variables. In the latter situation, the goodness of fit (GOF) criterion in multidimensional scaling-based linear calibration (MDSC) is also shown. In bold are the results corresponding to the MDS solution.

	MSEr of Ch		MSEr of MaxHR		
	Linear	Raking	Linear	Raking	
$\hat{\mu}_{yw}$	**0.9605031**	0.9970190	**0.8681743**	0.8650021	
r	$\hat{\mu}_{yw}^{MDS}(r)$	$\hat{\mu}_{yw}^{MDS}(r)$	$\hat{\mu}_{yw}^{MDS}(r)$	$\hat{\mu}_{yw}^{MDS}(r)$	GOF
1	0.9596628	0.9784779	0.7807146	0.7876983	0.3566573
2	0.9200081	0.9507770	0.7894168	0.7979797	0.4701502
3	0.8950116	0.9264501	0.8007014	0.8083127	0.5737942
4	0.9003786	0.9381284	0.8119167	0.8223844	0.6646301
5	0.9137757	0.9483126	0.8255316	0.8329744	0.7469115
6	0.9322494	0.9664986	***0.8336779***	***0.8451520***	***0.8146640***
7	0.9377027	0.9635124	0.8358269	0.8467278	0.8721036
8	***0.9313222***	***0.9569786***	0.8514425	0.8603327	***0.9189308***
9	0.9356105	0.9624573	0.8516795	0.8610817	0.9594853
10	0.9601692	0.9902532	0.8648146	0.8712115	0.9835898
11	0.9709255	1.0088332	0.8777309	0.8802210	1.000000

The proposed MDS-based procedure outperformed the traditional linear and raking calibration methods in all dimensions except the last one. To determine the number of MDS-based auxiliary variables we consider a value for the *GOF* coefficient above 80%, for which the lowest MESr value was found for $r = 6$ auxiliary variables in both distances for the *MaxHR* variable, which account for more than 81.5% of the variability. For the *Ch* variable, the lowest value was found for $r = 8$, accounting for more than 91.9% of the variability. Thus, the MDS procedure outperforms the traditional linear and raking calibration estimators whilst using fewer auxiliary variables. Furthermore and unlike the traditional calibration methods, for which linear and raking distances failed in approximately 1% of the samples generated, the MDS procedure always found the estimator value for the population mean.

Regarding the weights, which are common for both variables, the percentage of positive weights in each sample in the Monte Carlo procedure and the distribution of the coefficient of variation of the weights throughout the B samples were analysed. Up to dimension nine, the lowest percentage of positive weights in a sample exceeded 94% for the linear calibration procedure based on MDS, while for the last two dimensions, the proportion of positive weights exceeded 90%, reflecting the good performance of the proposed model. For both calibration procedures, high variability was observed with the original auxiliary variables, revealing distant outlier values. Thus, when performing traditional linear calibration, some values far from the coefficient of variation were obtained (45255.58 and 20817.53), and in general, more than 85.06% of the selected samples presented a coefficient greater than 0.3. The raking calibration procedure also produced high variability in the weights, such that in more than 69% of the samples generated, a coefficient of variation greater than 0.3 was observed. On the other hand, when the MDS-based auxiliary information was used, the linear calibration revealed low levels of dispersion in the estimated weights. Thus, for $r = 6$ and $r = 8$, only 4.06% and 9.4%, respectively, of the samples presented a coefficient of variation value greater than 0.3. Similar results were found for the raking calibration procedure, for which only 1.18% and 3.75% of the samples for the dimension $r = 6$ and $r = 8$, respectively, presented large dispersion values.

In summary, the results obtained show that the estimator of the mean can always be obtained with the MDS-based procedure. Moreover, it requires fewer auxiliary variables and produces a lower *MSEr*, compared with the traditional linear and raking calibration procedures on the original categorical variables, after recoding. When the linear calibration method was used in conjunction with the auxiliary variables derived from the MDS-based method, a high percentage of the weights obtained were positive. Moreover, the distribution of the weights for the Monte Carlo experiment presented less dispersion with the MDS-based procedure than with the original auxiliary information, with respect to both the linear and the raking calibration procedures.

5 Concluding remarks

The proposed calibration procedure has some advantages over other variable selection methods for calibration. First, it provides a single system of adjusted weights which does not depend on the target variables, a property that is very useful for multipurpose investigations such as health surveys, which often address several fields simultaneously and involve many variables on which specific analyses are based. It is very important in this type of survey to have a unique set of adjusted weights. Second, the procedure does not depend on the parameter to be estimated; instead, our calibration approach adapts to the estimation of more complex parameters than a population total, in several ways. One of the simplest is to use the substitution method, as follows: let θ_y be the parameter of interest that is a function of the total t_y, $\theta_y = f(t_y)$, thus a calibration estimator is given by $\hat{\theta} = f(\hat{t}_{yw}^{MDS}(r))$. If the parameter θ_y can be defined, as the solution to an implicit function known as the estimating equation (Godambe and Thompson[37]), a calibration estimator of the parameter of interest θ_y can be obtained as the solution to the estimating equation on s weighted by the calibration weights $w_k^{MDS}(r)$ in a similar way to the method described in Lesage,[38] section 3.2.

The method we propose can be used both for quantitative and for qualitative variables, in contrast to the use of principal component analysis,[28] which is only valid for quantitative variables. Furthermore, our proposal is computationally efficient, since the search procedure for the variables to be calibrated is done only once, whereas in the method proposed by Chauvet and Goga,[22] the variables to be calibrated must be selected for each variable objective, minimising the estimated variance by means of a bootstrap method, which increases the computational cost of the procedure.

Our procedure also has some drawbacks, chiefly the fact that, in general, the calibration property on the original variables is lost. The consistency of the weight system is an important issue for some users in official statistics but the goal of calibration is to decrease the variance of the H-T estimator.[9] The loss of the consistency property on the original variables can be assumed in some cases for the sake of efficiency in terms of reducing bias and errors. For example, the model-calibration technique proposed by Wu and Sitter[39] produces weights that when applied to the auxiliary variables, do not confirm known aggregates for these auxiliary variables (unless the model is linear). Penalised calibration[20,26,27] is other case which the resulting estimator does not necessarily have to be calibrated with respect to certain auxiliary variables, which makes the choice of the weights more flexible. Another approach that relaxes the calibration equations is to use principal component regression (Cardot et al.[28]).

6 Discussion

This article presents a procedure based on MDS to reduce the variance of the population estimator of the total or the mean of a variable, when using the survey calibration with auxiliary information that includes qualitative variables. The model produces a new set of uncorrelated continuous auxiliary variables, while preserving any existing linear relation between the target variable and the observed auxiliary variables, after the categorical variables are subjected to an appropriate scoring system.

By applying the all-purpose general similarity coefficient proposed by Gower,[33] a matrix of Euclidean distances is directly obtained from the observed qualitative or mixed auxiliary information, while the maximum number of dimensions of the metric MDS procedure does not exceed the total number of auxiliary variables after recoding.

The method we describe is based on projecting the target variable into a subspace spanned by the columns obtained by classical MDS, which play the role of predictive continuous independent auxiliary variables. Hence, the linear calibration procedure can be performed on this new set of auxiliary variables, which ensures the asymptotic design-consistency of this estimator. Since the auxiliary variables are associated with positive eigenvalues in the MDS procedure, the method always provides feasible estimators of the calibration weights.

In selecting the optimal number of auxiliary variables based on MDS, the following criteria were taken into consideration: a high percentage of variability explained by the MDS solution, a high proportion of estimated positive weights (for linear calibration), and a low variability in the distribution of the weights. In short, the number of new auxiliary variables was selected based on the lowest value of the dimensionality of the MDS solution consistent with an optimal combination of these properties.

The performance of the MDS-based calibration procedure is illustrated for real data sets involving qualitative auxiliary information, and the results obtained are also compared with those of traditional calibration procedures. In particular, the proposed procedure is compared with the traditional linear calibration method after qualitative variables are recoded into dummies. In general, the results obtained show that the proposed model always produces an estimated value of the mean, unlike the traditional calibration procedures, which in some situations within this framework may fail.

The proposed method was also analysed with each sample of a Monte Carlo procedure to estimate the mean for two different variables in which the total score in the population is known. This Monte Carlo experiment, too, was analysed with traditional linear and raking calibration procedures after recoding, and the *MSEr* was calculated. In general, the results obtained with the MDS-based procedure reflect a lower number of auxiliary variables, a lower *MSEr*, a higher proportion of positive weights, and a lower degree of dispersion in the distribution of the weights, with respect to both the linear and the raking calibration procedures.

The model we describe can be extended to other experimental situations. For example, an interesting case is when several categorical explanatory variables are combined to conform a large number of profiles.[40–42] In this situation, the combination of categories forming the profiles causes a large number of auxiliary variables to emerge after recoding using dummy variables, and a problem of over-calibration may arise in this qualitative information framework. Other methods for selecting the optimal set of variables are also being investigated, in particular, related to combined MDS and cluster methods that allow reducing dimensionality.[43,44]

A final consideration is that the use of linear calibration is motivated by the existence of a linear model. However, if the linear model does not fit and another type of curve relates the study variable with the auxiliary variables, an alternative model of calibrated estimators must be employed.[39,45] The application of the proposed methodology to the case of the calibration model is not immediate and will be studied in the future.

Acknowledgements

The authors wish to thank the Action Editor, and the anonymous reviewers whose comments greatly contributed to improving the quality of the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been partially supported by the Ministry of Science and Innovation – State Research Agency /10.13039/501100011033/ Spain, by 'ERDF A way of making Europe' [grant number RTI2018-099723-B-I00], and grant B-CTS-184-UGR20 funded by ERDF, EU / Ministry of Economic Transformation, Industry, Knowledge and Universities of Andalusia (J. Fernando Vera); MCIN/AEI/ 10.13039/501100011033: FEDER 'Una manera de hacer Europa' [grant number PID2019-106861RB-I00], and grant FEDER/Junta de Andalucía FQM170-UGR20 (M. Mar Rueda); and IMAG-Maria de Maeztu [grant number CEX2020-001105-M/AEI/10.13039/ 501100011033] (J. Fernando Vera, M. Mar Rueda).

ORCID iDs

J Fernando Vera https://orcid.org/0000-0002-6499-7132
Carmen Cecilia Sánchez Zuleta https://orcid.org/0000-0002-7410-9610
Maria del Mar Rueda https://orcid.org/0000-0002-2903-8745

Supplemental material

Supplemental material for this article is available online.

Note

1. The R script and dataset are available as supplementary material.

References

1. Pfeffermann D. The use of sampling weights for survey data analysis. *Stat Methods Med Res* 1996; **5**: 239-–261.
2. Deming WE and Stephan FF. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann Math Stat* 1940; **11**: 427–444.
3. Huang ET and Fuller WA. Nonnegative regression estimation for sample survey data. In: Proceedings of the Social Statistics Section; 1978. p. 300–305.
4. Deville JC and Särndal CE. Calibration estimators in survey sampling. *J Am Stat Assoc* 1992; **87**: 376–382.
5. Horvitz DG and Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 1952; **47**: 663–685.
6. Rueda M, Martínez S, Martínez H, et al. Estimation of the distribution function with calibration methods. *J Stat Plan Inference* 2007; **137**: 435–448.
7. Rueda M, Martínez-Puertas S and Martínez-Puertas Hea. Calibration methods for estimating quantiles. *Metrika* 2007; **66**: 355–371.
8. MdM Rueda. Comments on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *Test* 2019; **28**: 1077–1081.
9. Devaud D and Tillé Y. Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem. *Test* 2019; **28**: 1033–1065.
10. Särndal CE. The calibration approach in survey theory and practice. *Surv Methodol* 2007; **33**: 99–119.
11. Beaumont JF and Rao JNK. Comments on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *Test* 2019; **28**: 1071–1076.
12. Deville JC, Särndal CE and Sautory O. Generalized raking procedures in survey sampling. *J Am Stat Assoc* 1993; **88**: 1013–1020.

13. Zhang LC. Post-stratification and calibration-A synthesis. *Am Stat* 2000; **54**: 178–184.
14. Ranalli MG, Arcos A, Rueda MdM, et al. Calibration estimation in dual-frame surveys. *Stat Methods Appl* 2016; **25**: 321–349.
15. Wu C and Thompson ME. *Sampling Theory and Practice*. ICSA Book Series in Statistics. Cham: Springer International Publishing, 2020.
16. Kalton G and Flores-Cervantes I. Weighting methods. *J Off Stat* 2003; **19**: 81–97.
17. Estevao VM and Särndal CE. Survey estimates by calibration on complex auxiliary information. *International Statistical Review / Revue Internationale de Statistique* 2006; **74**: 127–147.
18. Detrano RC, Jánosi A, Steinbrunn W, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. *Am J Cardiol* 1989; **64**: 304–310.
19. Soriano FSEP. Heart Failure Prediction Dataset. Retrieved December 20; 2020. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.
20. Guggemos F and Tillé Y. Penalized calibration in survey sampling: design-based estimation assisted by mixed models. *J Stat Plan Inference* 2010; **140**: 3199–3212.
21. Nascimento Silva PLD and Skinner CJ. Variable selection for regression estimation in finite populations. *Surv Methodol* 1997; **23**: 23–32.
22. Chauvet G and Goga C. Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. *J Stat Plan Inference* 2022; **217**: 177–187.
23. Brick JM and Kalton G. Handling missing data in survey research. *Stat Methods Med Res* 1996; **5**: 215–238.
24. Clark RG and Chambers RL. Adaptive calibration for prediction of finite population totals. *Sur Methodol* 2008; **34**: 163–172.
25. Mcconville KS, Breidt Jay F, Lee TCM, et al. Model-assisted survey regression estimation with the lasso. *J Surv Stat Methodol* 2017; **5**: 131–158.
26. Beaumont JF and Bocci C. Another look at ridge calibration. *Metron* 2008; **66**: 5–20.
27. Barranco-Chamorro I, Jiménez-Gamero M, Mayor-Gallego JA, et al. A case-deletion diagnostic for penalized calibration estimators and BLUP under linear mixed models in survey sampling. *Comput Stat Data Anal* 2005; **87**: 18–33.
28. Cardot H, Goga C and Shehzad MA. Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Stat Sin* 2017; **27**: 243–260.
29. Borg I and Groenen PJF. *Modern Multidimensional Scaling. Theory and Applications*. New York: Second ed. Springer, 2005.
30. Särndal CE and Lundström S. *Estimation in surveys with nonresponse*. First ed. Hoboken, NY: John Wiley & Sons, Ltd, 2005.
31. Dubreuil G and Tremblay J. The use of generalized raking procedures to improve the quality of small domain estimation. In: *Statistics Canada*; 2002. p. 293–298.
32. Cuadras CM and Arenas C. A distance based regression model for prediction with mixed data. *Commun Stat - Theory Methods* 1990; **19**: 2261–2279.
33. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics* 1971; **27**: 857–871.
34. Vera JF and del Val EB. *Cluster distance-based regression*. In: Studies in Classification, Data Analysis, and Knowledge Organization. Springer Science and Business Media Deutschland GmbH; 2020. p. 385–397.
35. Mardia K, J B and Kent J. *Multivariate Analysis*. First ed. London: Academic Press, 1979.
36. Isaki CT and Fuller WA. Survey design under the regression superpopulation model. *J Am Stat Assoc* 1982; **77**: 89–96.
37. Godambe VP and Thompson ME. Parameters of superpopulation and survey population: their relationship and estimation. *Int Stat Rev* 1986; **54**: 127–138.
38. Lesage E. The use of estimating equations to perform a calibration on complex parameters. *Surv Methodol* 2011; **37**: 103–108.
39. Wu C and Sitter RR. A model-calibration approach to using complete auxiliary information from survey data. *J Am Stat Assoc* 2001; **96**: 185–193.
40. Vera JF, de Rooij M and Heiser WJ. A latent class distance association model for cross-classified data with a categorical response variable. *Br J Math Stat Psychol* 2014; **67**: 514–540.
41. Vera JF and De Rooij M. A latent block distance-association model for profile by profile cross-classified categorical data. *Multivariate Behav Res* 2020; **55**: 329–343.
42. Vera JF. Distance-based logistic model for cross-classified categorical data. *Br J Math Stat Psychol* 2022; **75**: 466–492.
43. Vera JF, Macías R and Heiser WJ. A latent class multidimensional scaling model for two-way one-mode continuous rating dissimilarity data. *Psychometrika* 2009; **74**: 297–315.
44. Vera JF and Macías R. On the behaviour of K-means clustering of a dissimilarity matrix by means of full multidimensional scaling. *Psychometrika* 2021; **86**: 489–513.
45. Rueda M, Sánchez-Borrego I, Arcos A, et al. Model-calibration estimation of the distribution function using nonparametric regression. *Metrika* 2009; **71**: 33–44.