Mathematics and Statistics 10(1): 100-115, 2022
DOI: 10.13189/ms.2022.100108 http://www.hrpub.org

A Goal Programming Approach for Generalized Calibration Weights Estimation in Stratified Random Sampling

Siham Rabee*, Ramadan Hamed, Ragaa Kassem, Mahmoud Rashwaan

Department of Statistics, Faculty of Economics and Political Science, Cairo University, Egypt

Received October 1, 2021; Revised November 28, 2021; Accepted December 22, 2021

Cite This Paper in the following Citation Styles

(a): *[1] Siham Rabee, Ramadan Hamed, Ragaa Kassem, Mahmoud Rashwaan , "A Goal Programming Approach for Generalized Calibration Weights Estimation in Stratified Random Sampling," Mathematics and Statistics, Vol. 10, No. 1, pp. 100 - 115, 2022. DOI: 10.13189/ms.2022.100108.*

(b): *Siham Rabee, Ramadan Hamed, Ragaa Kassem, Mahmoud Rashwaan (2022). A Goal Programming Approach for Generalized Calibration Weights Estimation in Stratified Random Sampling. Mathematics and Statistics, 10(1), 100 - 115. DOI: 10.13189/ms.2022.100108.*

Abstract Calibration estimation approach is a widely used method for increasing the precision of the estimates of population parameters. It works by modifying the design weights as little as possible by minimizing a given distance function to the calibrated weights respecting a set of constraints related to specified auxiliary variables. This paper proposes a goal programming approach for generalized calibration estimation. In the generalized calibration estimation, multi study variables will be considered by incorporating multi auxiliary variables. Almost all calibration estimation's literature proposed calibrated estimators for the population mean of only one study variable. And nevertheless, up to researcher's knowledge, there is no study that considers calibration estimation approach for multi study variables. According to the correlation structure between the study variables, estimating the calibrated weights will be formulated in two different models. The theory of the proposed approach is presented and the calibrated weights are estimated. A simulation study is conducted in order to evaluate the performance of the proposed approach in the different scenarios compared by some existing calibration estimators. The Simulation results of the four generated populations show that the proposed approach is more flexible and efficient compared to classical methods.

Keywords Generalized Calibration Estimation, Multivariate Study Variables, Auxiliary Variables, Stratified Sampling, Goal Programming Technique

1. Introduction

Calibration estimation provides the statisticians a chance for improving the estimates of the population parameters by incorporation the auxiliary information in the estimation procedure. By the auxiliary information we mean the additional variables that are correlated in some way with the study variable. The auxiliary information can be available for the target population, either in an aggregated form or in a detailed form for each individual population unit. So the main purpose behind MCE is to derive calibrated weights that modify as little as possible the design weights which have the desirable property of yielding unbiased estimates. Accordingly, the survey statistician wants to stay as close as possible to these design weights.

Deville and Sarndal [1] first introduced calibration estimation in survey sampling. Since then, many researchers have contributed in calibration estimation approach to develop calibrated estimators for different population parameters using different constraints under different sampling schemes. Singh [2] is considered the first contributor that extended the calibration approach to Stratified Random Sampling (SRS) design. Following [2],

many authors have contributed to calibration estimation's theory in SRS [3-5].

This paper can be considered as an extension for a paper introduced by [6]. Accordingly, this paper proposes a goal programming approach for Generalized Calibration Estimation (GCE). This new approach will focus on calibration estimation for the population means of multi study variables under SRS scheme by incorporating multi auxiliary variables. Almost all calibration estimation's literature proposed calibrated estimators for the population mean of only one study variable. And nevertheless, up to researcher's knowledge, there is no study that considers calibration estimation approach for multi study variables.

This paper formulates the generalized approach of the calibration estimation weights as a Mathematical Programming Problem (MPP) in which the Manhattan distance is minimized subject to some important constraints. In addition to the calibration constraints, some constraints that improve the properties of the calibrated estimators should be taken into consideration. The considered optimization problem will be solved using Goal Programming technique. The computational details will be illustrated using a simulation study considering two study variables and two auxiliary variables as a special case of the generalized calibration estimation approach.

The remaining part of this paper is organized as follows; Section 2 presents notations about calibration approach in SRS scheme. The literature of the calibration estimation approach is presented in Section 3. Section 4 deduces a MPP for GCE in the case of ignoring the correlation between the study variables along with its suitable solving method. Section 5 presents the MPP for the proposed approach of the calibration estimation in which the correlation between the study variables is considered. In order to evaluate the performance of the suggested approach through a simulation study, two study variables and two auxiliary variables are considered as a special case of the GCE approach in section 6. Finally, section 7 summarizes the main conclusions and contribution of this paper.

2. Basic Notations of Calibration Estimation Approach in Stratified Random Sampling

Consider a finite population sized N units with L homogeneous subgroups called strata. Each stratum's size is N_h units э $\sum_{h=1}^{L} N_h = N$; h = 1, 2 . . . L and $W_h = N_h / N$ is the stratum's weight. For this population, a sample of size n_h is drawn by Simple Random Sampling Without Replacement (SRSWOR) from each stratum э $\sum_{h=1}^{L} n_h = n$, where n is the total sample size. Suppose the i^{th} unit's value of the j^{th} study variable selected from the h^{th} stratum is denoted by y_{hji}; j=1, 2 & i=1, 2…. n_h. It must be noted that the unbiased estimator of the population mean of the j^{th} study variable under SRS design $\overline{Y}_j = \sum_{h=1}^{L} W_h \overline{Y}_{hj}$ is given by [7]

$$\overline{y}_{j\,(st)} = \sum_{h=1}^{L} W_h \overline{y}_{hj} \qquad (2.1)$$

where $\overline{y}_{hj} = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hji}$ denotes the h^{th} stratum sample mean. Additionally, the estimated variance of $\overline{y}_{j\,(st)}$ under the SRSWOR scheme is given by [7]

$$\hat{v}\,(\overline{y}_{j(st)}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy_j}^2 \qquad (2.2)$$

where $s_{hy_j}^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hji} - \overline{y}_{hj})^2$ is the sample variance of Y_j in stratum h, and $f_h = n_h / N_h$ is the sampling fraction for the h^{th} stratum; h=1,2,…L & j=1, 2.

Assume that X_{hji} denotes the i^{th} unit's value of the j^{th} auxiliary variable in the stratum h; h=1,2,….L, $i = 1,2,\ldots.$, N_h and j=1,2. The stratum means $\overline{X}_{hj} = \frac{1}{N_h}\sum_{i=1}^{Nh} x_{hji}$ and the stratum variances $S^2{}_{hj} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(x_{hji} - \overline{X}_{hj})^2$ of each auxiliary variable are assumed to be known. Since the main purpose of any calibration estimation approach is to introduce an improved estimator of the population's parameter by incorporating auxiliary information. Accordingly, the population mean's calibration estimator, under SRS, can be given by [3]

$$\overline{y}_j^{\,c} = \sum_{h=1}^{L} \Omega_h \overline{y}_{hj}\ , \qquad (2.3)$$

where Ω_h ; h=1,2….L denote the estimated calibration weights that are chosen to minimize a given distance function, subject to some specific calibration constraints. In addition, the calibrated estimator's variance can be computed by substituting the design weights, W_h with the calibration weights Ω_h; h=1, 2…L in (2.2). Accordingly, the calibrated estimator's variance can be given by [3]

$$\hat{v}(\overline{y}_j^{\,c}) = \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hyj}^2\ , \qquad (2.4)$$

3. Review for the Previous Calibration Estimators in Stratified Random Sampling

Using Mathematical programming tools in calibration estimation is suggested in this paper since they offer researchers the advantages of optimizing an objective function with respect to a set of system constraints identifying the main problem's structure. This is one of the main features that have been utilized by many authors in the field of calibration estimation. This section presents some mathematical programming models, which have been suggested in the literature to determine the optimal calibrated weights.

The calibration estimation's literature can be classified into two groups based on the number of incorporated auxiliary variables; the first group represents the studies that contributed in univarite calibration estimation by incorporating one auxiliary variable, while the second group introduces some multivariate calibration estimators

using two auxiliary variables.

The calibration estimator introduced by Singh [3] is considered one of the main estimators for the population mean under SRS scheme classified in the first group. Singh [3] minimized the chi-square distance function subject to two basic calibration constraints. In the first constraint it is assumed that the weighted sum of the sample's mean of the auxiliary variable has to be equal to the known parameter of that auxiliary variable. While the second constraint indicates that the sum of the calibrated weights is equal to 1. So Singh [3] optimization problem was formulated as follows

Find Ω_h that

$$\text{Minimize } \sum_{h=1}^{L} \frac{(W_h-\Omega_h)^2}{W_h Q_h} \quad (3.1)$$

Subject to

$$\sum_{h=1}^{L} \Omega_h \bar{x}_h = \bar{X} \quad , \quad (3.2)$$

$$\sum_{h=1}^{L} \Omega_h = 1 \quad (3.3)$$

where $\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$ and $\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$ are the auxiliary variable's sample and population means in h^{th} stratum respectively. Q_h are appropriately chosen constant that determine the final form of the calibrated estimators.

Moreover, Tracy [4] introduced their calibration estimator by minimizing the chi-square distance function subject to two constraints. The first one is the same as the constraint in (3.2), while the second constraint used the second order moments of the auxiliary variable. Hence, the second constraint can be expressed as follows:

$$\sum_{h=1}^{L} \Omega_h {s^2}_h = S^2, \quad (3.4)$$

where s_h^2 and $S^2 = \sum_{h=1}^{L} W_h {S^2}_h$ are the auxiliary variable's sample and population variance in the h^{th} stratum respectively.

Motivated by Singh [3] and Tracy [4], Koyuncu and Khadilar [8] proposed their calibration estimator. They also minimized the chi-square distance function subject to three constraints which are the same as the constraints proposed in [3] and in [4]. But they considered using them at one optimization problem simultaneously. Accordingly it can be stated that they minimized (3.1) subject to the three constraints expressed in (3.2), (3.3) and (3.4) together simultaneously [8].

Regarding the second group that focused on MCE using two auxiliary variables, Rao [5] can be considered the first study that proposed incorporating two auxiliary variables in MCE for the population mean. They considered minimizing (3.1) subject to one constraint in which the weighted sum of the j^{th} auxiliary variable's sample means is equal to the known parameter for that auxiliary variable. Their optimization problem is expressed by

Find Ω_h that

$$\text{Minimize} \sum_{h=1}^{L} \frac{(W_h-\Omega_h)^2}{W_h Q_h} \quad (3.1)$$

Subject to

$$\sum_{h=1}^{L} \Omega_h \bar{x}_{hj} = \bar{X}_j \quad , \quad (3.5)$$

where $\bar{x}_{hj} = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hji}$ and $\bar{X}_j = \sum_{h=1}^{L} W_h \bar{X}_{hj}$ are the j^{th} auxiliary variable's sample and population means in h^{th} stratum respectively.

Further, Rao [9] proposed a MCE for the population mean in SRS using different constraints. They also suggested minimizing (3.1) as their objective function subject to the same three constraints used in [8] but in case of multi-auxiliary variables.

Finally, the new approach for MCE proposed by [6] is considered a recent study that differs from all the previously mentioned literature. While almost all calibration estimation's literature used the Lagrange multiplier technique in order to estimate the calibrated weights, goal programming approach was suggested by [6] to solve their MMP. Moreover, the Manhattan distance (L1 Norm) from the design weights W_h is minimized as their main objective function. So their optimization problem is expressed by

Find Ω_h that

$$Min\ Z = \sum_{h=1}^{L} |W_h - \Omega_h|, \quad (3.6)$$

Subject to

$$\sum_{h=1}^{L} \Omega_h \bar{x}_{hj} = \bar{X}_j, \quad j=1, 2 \quad (3.7)$$

$$\sum_{h=1}^{L} \Omega_h {s^2}_{hj} = {S^2}_j, \quad j=1, 2 \quad (3.8)$$

$$\sum_{h=1}^{L} \Omega_h = 1 \ , \quad (3.9)$$

$$\sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy}^2 \ \leq \ \zeta, \quad (3.10)$$

$$\left|\sum_{h=1}^{L} \Omega_h \bar{y}_h - \bar{Y}\right| \leq \epsilon, \quad (3.11)$$

$$\Omega_h \geq 0, \ h=1, 2, .., L \quad (3.12)$$

where W_h & Ω_h are the design and calibrated weights, respectively; $h=1, 2, .., L$.

$\sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy}^2$ denotes the estimated variance of the population parameter's calibrated estimator.

ζ is a positive upper bound for the estimated variance of the calibrated estimator.

$\left|\sum_{h=1}^{L} \Omega_h \bar{y}_h - \bar{Y}\right|$ denotes the estimated bias of the population parameter's calibrated estimator.

ϵ is a specified positive constant as an upper bound for the estimated bias of the calibrated estimator

Almost all previously mentioned literature proposed calibrated estimators for the population mean of one study variable only. And nevertheless, up to the researcher's knowledge, there is no study that considered calibration estimation approach for multi study variables. So this paper will focus on generalized calibration estimation for the population mean of multi study variables simultaneously incorporating multi auxiliary variables.

It is expected that the GCE for the population parameters of multi study variables, simultaneously, may differ

according to the correlation structure between these study variables. Accordingly, this paper suggests formulating the problem of GCE in two different MP models. The first one that represents the Mathematical Programming Problem (MPP) for calibration estimation when the correlation between the study variables is ignored, while the second model represents the case in which the correlation between the study variables is considered.

4 The Suggested GCE Approach Ignoring the Correlation between the Study Variables

This section presents a Mathematical Programming model for generalized calibration estimation when the correlation structure between the study variables is ignored. The decision variables in the suggested model are the L calibrated weights Ω_h ; h=1, 2…L assuming that the population under the study consists of L strata. The objective function and the constraints of the proposed model will be presented in subsections (4.1) and (4.2). Finally the proposed MP model will be presented in subsection (4.3) along with its suitable solving method in subsection (4.4).

4.1. The Objective Function:

This model is concerned with minimizing the same objective function considered in (3.6), that represents a Manhattan distance measure between the design weights (W_h ; h=1…L) and the calibrated weights (Ω_h ; h=1…L). So it can be expressed as follows:

$$\text{Minimize } Z=\sum_{h=1}^{L} | W_h - \Omega_h | \qquad (4.1)$$

where W_h and Ω_h are the design weight and calibrated weight for the h[th] stratum respectively; h=1, 2…L.

4.2. The Constraints

Regarding the constraints considered in this model, they are the same constraints presented previously in [6] expressed in (3.7 - 3.12) but with modifying them in the light of having multi study variables. The first set represents the calibration constraints and the second set contains the constraints that concentrate on improving the precision of the calibrated estimators. With respect to the first set "the calibration constraints set" that depends on incorporating the auxiliary variables is the same set considered in [6] without modification.

In the second set of the constraints that concentrate on improving the precision of the calibrated estimators, there are two constraints included in the MMP model. The first constraint is designed to guarantee that the estimated variance of the calibrated estimators is less than or equal to a specified positive constant ζ_j, and can be expressed as follows:

$$\sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{yhj}^2 \le \zeta_j, \; j=1,2,\ldots p \qquad (4.2)$$

where $s^2_{yhj} = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hji} - \overline{y}_{hj})$ is the h[th] stratum sample variance of the j[th] study variable Y_j, and $f_h = n_h / N_h$ is the sampling fraction for the h[th] stratum; h=1,2,…L.

Moreover, the second constraint is designed to guarantee that the improved calibrated estimator is still close to unbiasness. Hence, it can be expressed as follows

$$\left| \sum_{h=1}^{L} \Omega_h \overline{y}_{hj} - \overline{Y}_j \right| \le \epsilon_j, \; j=1,2,\ldots p \qquad (4.3)$$

where $\sum_{h=1}^{L} \Omega_h \overline{y}_{hj}$ represents the calibration estimator of the population mean of the j[th] study variable, and $\overline{Y}_j$ denotes its parameter, ϵ_j is a specified positive constant.

4.3. The MMP Model

The main structure for the MMP model can be expressed according to the model's elements presented in the previous subsections as follows:

Find Ω_h that

$$Min\ Z = \sum_{h=1}^{L} | W_h - \Omega_h | \qquad (4.4)$$

Subject to

$$\sum_{h=1}^{L} \Omega_h \overline{x}_{hj} = \overline{X}_j, \; j=1, 2\ldots p \qquad (4.5)$$

$$\sum_{h=1}^{L} \Omega_h s^2_{hj} = S^2_j, \; j=1, 2\ldots p \qquad (4.6)$$

$$\sum_{h=1}^{L} \Omega_h = 1 \qquad (4.7)$$

$$\sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s^2_{hyj} \le \zeta_j, \; j=1, 2\ldots p \qquad (4.8)$$

$$\left| \sum_{h=1}^{L} \Omega_h \overline{y}_{hj} - \overline{Y}_j \right| \le \epsilon_j, \; j=1, 2\ldots p \qquad (4.9)$$

$$\Omega_h \ge 0, \; h=1, 2...L$$

where;

W_h & Ω_h is the design and calibrated weights, respectively; *h=1, 2, .. , L.*

$\overline{x}_{hj}$ is sample mean for the j[th] auxiliary variable in h[th] stratum; *h=1, 2, .. , L.*

$\overline{X}_j$ is the population mean for the j[th] auxiliary variable.

s^2_{hj} is sample variance for the j[th] auxiliary variable in h[th] stratum; h=1,2…L.

S^2_j is the population variance for the j[th] auxiliary variable.

$\sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s^2_{hy_j}$ denotes the estimated variance of the population mean's calibrated estimator for the j[th] study variable.

ζ_j is a positive upper bound for the estimated variance of the calibrated estimator.

$\left| \sum_{h=1}^{L} \Omega_h \overline{y}_{hj} - \overline{Y}_j \right|$ denotes the estimated biasness of the population mean 's calibrated estimator for j[th] study variable.

ϵ_j is a specified positive constant as an upper bound for the estimated bias of the calibrated estimator

4.4. The GP Model

The considered optimization problem, in subsection 4.3, is suggested to be solved using Goal Programming technique by adding positive and negative deviational variables to the main objective function. Since it can be proved that minimizing the absolute deviation between any parameter and its estimate is equivalent to minimizing the sum of the positive and negative deviational variables added to this function [10]. This practically means that minimizing the sum of positive and negative deviational variables for any objective (goal constraint) is equivalent to minimizing the absolute deviation between this goal constraint and its target value. Hereafter there is a mathematical proof that can be used to clarify the previous statement not only for the two-variable case, but also for more general multivariate conditions.

Considering the objective (goal constraint) expressed in (4.10) as follows:

$$aX_1 + bX_2 + n_1 - p_1 = \theta, \quad (4.10)$$

where θ is the target value for this objective (goal constraint), $n_1 \& p_1$ are called positive and negative deviational variable respectively.

The positive and negative deviational variables can be defined by the following [10]:

$$\boldsymbol{n_1} = \frac{1}{2}\,[\,\theta - aX_1 + bX_2 + |\,\theta - aX_1 + bX_2\,|] \quad (4.11)$$

$$\boldsymbol{p_1} = \frac{1}{2}\,[aX_1 + bX_2 - \theta + |\,\theta - aX_1 + bX_2\,|] \quad (4.12)$$

Accordingly, from (4.11) & (4.12), we get:

$$\boldsymbol{n_1 + p_1} = \frac{1}{2}\theta - \frac{1}{2}(aX_1 + bX_2) + \frac{1}{2}|\,\theta - aX_1 + bX_2\,| + \frac{1}{2}(aX_1 + bX_2) - \frac{1}{2}\theta + \frac{1}{2}|\,\theta - aX_1 + bX_2\,|.$$

Hence $n_1 + p_1 = |\,\theta - aX_1 + bX_2\,|$.

Through applying the previously proofed procedure and dividing each objective (goal constraint) by its target value; the suggested MMP model can be converted to the following GP model:

Find Ω_h, dn_{1h}, dp_{1h} ; h=1,2..L. dn_k, dp_k; k=2, 3,..9, that

$$\text{Min } Z = \sum_{h=1}^{L}(dn_{1h} + dp_{1h}) + dp2 + dp3 + \sum_{k=4}^{9}(dn_k + dp_k) \quad (4.13)$$

subject to;

$$[\Omega_h/W_h] + dn_{1h} - dp_{1h} = 1;\ h=1, 2,.\ L \quad (4.14)$$

$$[\sum_{h=1}^{L} \Omega_h^2 (\tfrac{1-f_h}{n_h})\, s_{hy1}^2]/\zeta_1 + dn_2 - dp_2 = 1; \quad (4.15)$$

$$[\sum_{h=1}^{L} \Omega_h^2 (\tfrac{1-f_h}{n_h})\, s_{hy2}^2]/\zeta_2 + dn_3 - dp_3 = 1; \quad (4.16)$$

$$[\sum_{h=1}^{L} \Omega_h \bar{x}_{h1}]/\,\overline{X_1} + dn_4 - dp_4 = 1; \quad (4.17)$$

$$[\sum_{h=1}^{L} \Omega_h \bar{x}_{h2}]\,/\,\overline{X_2} + dn_5 - dp_5 = 1; \quad (4.18)$$

$$[\sum_{h=1}^{L} \Omega_h s^2{}_{1h}]\,/S^2{}_1 + dn_6 - dp_6 = 1; \quad (4.19)$$

$$[\sum_{h=1}^{L} \Omega_h s^2{}_{2h}]\,/\,S^2{}_2 + dn_7 - dp_7 = 1; \quad (4.20)$$

$$[\sum_{h=1}^{L} \Omega_h \bar{y}_{h1}]\,/\,\overline{Y_1} + dn_8 - dp_8 = 1; \quad (4.21)$$

$$[\sum_{h=1}^{L} \Omega_h \bar{y}_{h2}]\,/\,\overline{Y_2} + dn_9 - dp_9 = 1; \quad (4.22)$$

$$\sum_{h=1}^{L} \Omega_h = 1; \quad (4.23)$$

Ω_h , dn_{1h}, dp_{1h} , h=1,2..L. dn_k , dp_k ; ≥ 0; k=2, 3 ..9

Since there is a theoretical negative relation between the variance and the bias of any estimator, trying to decrease the variance of the estimators may cause getting biased estimators. So it can be said that there is a tradeoff between decreasing both of the variance and the bias of the calibrated estimator. Accordingly; three cases for the proposed GP approach will be considered. The main difference between these cases depends on the relative importance given to both of the variance and bias goal constraints. The relative importance of any goal constraint can be determined by the priority weights given to the deviational variables added to each goal constraint.

The three cases can be further explained as follows:

Case 1: Both of the variance and the bias of each calibrated estimator has the same relative importance so that equal priority weights are given to both dp_2 and $(dn_8 + dp_8)$ deviational variables of the variance and the bias goal constraints for the calibrated estimator of $\overline{Y_1}$. Similarly for the calibrated estimator of $\overline{Y_2}$; equal priority weights are given to both dp_3 and $(dn_9 + dp_9)$ deviational variables.

Case 2: In case 2, the variance of each calibrated estimator is considered relatively more important than its unbiasness. Hence the priority weight given to dp_2 is equal to 1, while the priority weight for $(dn_8 + dp_8)$ is zero, and the priority weight given to dp_3 is equal to 1, while the priority weight for $(dn_9 + dp_9)$ is zero.

Case 3: this case can be considered the inverse of case 2, i.e. getting an unbiased calibrated estimator is more important than decreasing its variance. So the priority weight given to dp_2 equals to zero, while the priority weight for $(dn_8 + dp_8)$ is one. Similarly for the calibrated estimator of $\overline{Y_2}$ the priority weight given to dp_3 equals to zero, while the priority weight for $(dn_9 + dp_9)$ is one.

5. The Suggested GCE Approach Considering the Correlation between the Two Study Variables

This section presents a Mathematical Programming model for the Generalized Calibration Estimation (GCE) in which the correlation structure between the two study variables is considered. Like the previous model; the decision variables in the suggested model are the L calibrated weights Ω_h ; h=1, 2…L assuming that the population under study consists of L strata. The following subsections give a thorough explanation of the suggested objective function and the constraints of the proposed model in subsections (5.1) and (5.2) respectively; finally

the proposed MP model will be presented in subsection (5.3) along with its suitable solving method in subsection (5.4).

5.1. The Objective Function

This model is concerned with minimizing the same objective function considered in subsection 4.1, that represents a Manhattan distance measure between the design weights (W_h ; h=1…L) and the calibrated weights (Ω_h ; h=1…L).

5.2. The Constraints

This model includes the same constraints previously detailed in subsection 4.2 with some modification in the variance constraints. The modification is suggested in the light of considering the correlation structure between the two study variables in the model. So that the same two types of constraints are included, the first set represents the calibration constraints and the second contains the constraints that concentrate on improving the precision of the calibrated estimators.

Regarding the variance constraint's modification, the correlation structure between the multi study variables is considered. So the variances of the calibrated estimators are presented as variance covariance matrix. The variance covariance matrix is a symmetric matrix in which the elements on the main diagonal represent the estimated variances of the calibrated estimators, while the off diagonal's elements represent the estimated covariance between each pair of calibrated estimators. So $cov\ (\overline{y}_{1st}^{\,c}, \overline{y}_{2st}^{\,c})$ can be expressed by the following matrix:

$$\begin{bmatrix} \hat{v}(\overline{y}_1^{\,c}) & \cdots & \widehat{cov}(\overline{y}_1^{\,c}, \overline{y}_p^{\,c}) \\ \vdots & \ddots & \vdots \\ \widehat{cov}(\overline{y}_1^{\,c}, \overline{y}_p^{\,c}) & \cdots & \hat{v}(\overline{y}_p^{\,c}) \end{bmatrix} \quad (5.1)$$

where,

$$\hat{v}(\overline{y}_1^{\,c}) = \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy_1}^2 \quad (5.2)$$

$$\hat{v}(\overline{y}_p^{\,c}) = \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy_p}^2 \quad (5.3)$$

Since the estimated covariance between each pair of the calibrated estimators can be given by [11] as follows:

$$\widehat{cov}\left(\overline{y}_i^{\,c}, \overline{y}_j^{\,c}\right) = \sum_{h=1}^{L} {W_h^2}/{n_h}\ \widehat{cov}_h(y_1, y_p) - \sum_{h=1}^{L} {W_h}/{n_h}\ \widehat{cov}_h(y_1, y_p) \quad (5.4)$$

$$\widehat{cov}\left(\overline{y}_1^{\,c}, \overline{y}_p^{\,c}\right) = \sum_{h=1}^{L} {W_h^2}/{n_h}\,(1 - {1}/{n_h})\ \widehat{cov}_h(y_1, y_p) \quad (5.5)$$

Accordingly, the included constraint to guarantee that the estimated variances of the calibrated estimators are less than or equal to a positive upper bound can be expressed as follows:

$$\begin{bmatrix} \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy1}^2 & \cdots & \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) \mathrm{cov}_h(y_1, y_p) \\ \vdots & \ddots & \vdots \\ \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) \mathrm{cov}_h(y_1, y_p) & \cdots & \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy_p}^2 \end{bmatrix} \leq \zeta \quad (5.6)$$

Such that

$$\zeta = \begin{bmatrix} \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy1}^2 & \cdots & \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) \mathrm{cov}_h(y_1, y_p) \\ \vdots & \ddots & \vdots \\ \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) \mathrm{cov}_h(y_1, y_p) & \cdots & \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy_p}^2 \end{bmatrix} \quad (5.7)$$

Moreover the constraints, designed to guarantee that the improved calibrated estimators are still close to be unbiased, can be expressed as follows

$$\left|\sum_{h=1}^{L} \Omega_h \overline{y}_{hj} - \overline{Y}_j\right| \leq \epsilon_j;\ j=1, 2...p \quad (5.8)$$

5.3. The MMP Model

The main structure for the MMP model in the case of considering the correlation between the study variables can be expressed according to its elements presented in the previous subsections as follows:

Find Ω_h that

$$\text{Min } Z = \sum_{h=1}^{L} |W_h - \Omega_h| \quad (5.9)$$

Subject to

$$\sum_{h=1}^{L} \Omega_h \overline{x}_{hj} = \overline{X_j},\ \ j=1, 2..p \quad (5.10)$$

$$\sum_{h=1}^{L} \Omega_h s^2{}_{hj} = S^2{}_j\ \ j=1, 2..p \quad (5.11)$$

$$\sum_{h=1}^{L} \Omega_h = 1 \quad (5.12)$$

$$\begin{bmatrix} \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy_1}^2 & \cdots & \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) \mathrm{cov}_h(y_1, y_p) \\ \vdots & \ddots & \vdots \\ \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) \mathrm{cov}_h(y_1, y_p) & \cdots & \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy_p}^2 \end{bmatrix} \leq \zeta \quad (5.13)$$

$$\left|\sum_{h=1}^{L} \Omega_h \overline{y}_{h1} - \overline{Y_1}\right| \leq \epsilon_1, \quad (5.14)$$

$$\left|\sum_{h=1}^{L} \Omega_h \overline{y}_{h2} - \overline{Y_2}\right| \leq \epsilon_2, \quad (5.15)$$

$$\Omega_h \geq 0,\ h=1, 2...\ , L$$

5.4. The GP Model

The considered MMP in (5.9 - 5.15) is suggested to be solved using Goal Programming technique. The computational details of the proposed calibration estimation approach will be illustrated for two study variables and by incorporating two auxiliary variables as the MMP in subsection 5.3. The same procedure used previously, to convert the MMP in 5.3 into a GP model, is used here also. With respect to converting the constraint (5.13) into a goal constraint, it is suggested to get the determinant function of the matrices in both sides (Left hand side and its upper bound in the Right hand side). So the constraint in (5.13) can be converted to the determinant function as follows:

$$\left|\left\{\sum_{h=1}^{L}\Omega_h^2\left(\frac{1-f_h}{n_h}\right)s_{hy1}^2 * \sum_{h=1}^{L}\Omega_h^2\left(\frac{1-f_h}{n_h}\right)s_{hy2}^2\right\} - \left\{\sum_{h=1}^{L}\Omega_h^2\left(\frac{1-f_h}{n_h}\right)\mathrm{Cov}_h(y_1,y_2) * \sum_{h=1}^{L}\Omega_h^2\left(\frac{1-f_h}{n_h}\right)\mathrm{Cov}_h(y_1,y_2)\right\}\right| \le |\zeta| \quad (5.16)$$

where

$$|\zeta| = \left|\left\{\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)s_{hy1}^2 * \sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)s_{hy2}^2\right\} - \left\{\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)\mathrm{Cov}_h(y_1,y_2) * \sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)\mathrm{Cov}_h(y_1,y_2)\right\}\right| \quad (5.17)$$

Accordingly; the suggested goal programming model can be expressed as follows:

Find Ω_h, dn_{1h}, dp_{1h} ; h=1,2,..L. dn_k, dp_k; k=2, 3,..8, that

$$\text{Min } Z = \sum_{h=1}^{L}(dn_{1h} + dp_{1h}) + dp2 + dp3 + \sum_{k=4}^{8}(dn_k + dp_k) \quad (5.18)$$

subject to;

$$[\Omega_h / W_h] + dn_{1h} - dp_{1h} = 1;\ h=1, 2,.\ L \quad (5.19)$$

$$\left(\left|\left\{\sum_{h=1}^{L}\Omega_h^2\left(\frac{1-f_h}{n_h}\right)s_{hy1}^2 * \sum_{h=1}^{L}\Omega_h^2\left(\frac{1-f_h}{n_h}\right)s_{hy2}^2\right\} - \left\{\sum_{h=1}^{L}\Omega_h^2\left(\frac{1-f_h}{n_h}\right)\mathrm{Cov}_h(y_1,y_2) * \sum_{h=1}^{L}\Omega_h^2\left(\frac{1-f_h}{n_h}\right)\mathrm{Cov}_h(y_1,y_2)\right\}\right| / |\zeta|\right) + dn_2 - dp_2 = 1 \quad (5.20)$$

$$\left[\sum_{h=1}^{L}\Omega_h\bar{x}_{h1}\right] / \bar{X}_1 + dn_3 - dp_3 = 1 \quad (5.21)$$

$$\left[\sum_{h=1}^{L}\Omega_h\bar{x}_{h2}\right] / \bar{X}_2 + dn_4 - dp_4 = 1; \quad (5.22)$$

$$\left[\sum_{h=1}^{L}\Omega_h s^2{}_{1h}\right] / S^2{}_1 + dn_5 - dp_5 = 1; \quad (5.23)$$

$$\left[\sum_{h=1}^{L}\Omega_h s^2{}_{2h}\right] / S^2{}_2 + dn_6 - dp_6 = 1; \quad (5.24)$$

$$\left[\sum_{h=1}^{L}\Omega_h\bar{y}_{h1}\right] / \bar{Y}_1 + dn_7 - dp_7 = 1; \quad (5.25)$$

$$\left[\sum_{h=1}^{L}\Omega_h\bar{y}_{h2}\right] / \bar{Y}_2 + dn_8 - dp_8 = 1; \quad (5.26)$$

$$\sum_{h=1}^{L}\Omega_h = 1; \quad (5.27)$$

Ω_h , dn_{1h}, dp_{1h} , h=1,2...L. dn_k , dp_k ; ≥ 0; k=2, 3, ..8

6. Simulation Study

In this section, a simulation study is conducted to assess the performance of the proposed approach for generalized calibration estimation for the population mean of two study variables. It must be noted that providing a simulation results for more than two study variables requires a fundamental change in the proposed model resulting a considerable theoretical and computational difficulties. Accordingly, the simulation study is conducted for two study and auxiliary variables as a special case for simplicity.

The Relative Root Mean Square Error (RRMSE %) criteria will be used for assessing the performances of different calibration estimators. It can be defined by [12] as follows:

$$\text{RRMSE} = \sqrt{\frac{\sum_{K=1}^{1000}\left(\frac{\bar{y}_{st}^{(\alpha)} - \bar{Y}}{\bar{Y}}\right)^2}{1000}} \quad (6.1)$$

where $\bar{Y}$ denotes the population parameter and $\bar{y}^{(\alpha)}$ denotes the calibration estimators. In addition to comparing the proposed models with some models presented in the literature. The following subsections present the conducted simulation study.

6.1. The Design of the Simulation Study

Throughout this study, the aim is to explore the effect of the following factors on the performance of the proposed models in the following different scenarios:

1. The probability distribution from which both of the study and auxiliary variables are generated. Accordingly, four different populations will be considered.
2. The correlation structure between the two study variables. Two scenarios will be studied, the first when considering the correlation between the study variables and the second when ignoring it.
3. The relative importance of both of the variance and biasness of the calibrated estimators, which is determined by the priority weights given to the deviational variables of each goal constraint. Three cases, detailed previously in subsection 4.4, will be considered.

The fixed factors in the simulation study are: the population size N, the number of strata and the number of both of the study and auxiliary variables. Moreover different scenarios are considered to assess the performance of the proposed approach for GCE.

6.2. Data Generation

In this simulation study, four different populations are generated from different distributions. Each population consists of three strata within each stratum the generated variables followed the distributions depicted in Table 1. The strata's sizes are taken as 3000, 2000, 5000 respectively. For each population, 1000 datasets are generated where each dataset's size is 10,000 observations.

Moreover, a 1000 sized sample is proportionally selected from each population under stratified sampling scheme. For each population and its selected sample, the means and variances for the four generated variables are calculated within the three strata. These calculated values for the means and the variances represent the input for the

models under comparison.

With respect to the properties of the four generated populations; all variables are distributed normally in each stratum for the first population, while all variables are distributed skewed positively in each stratum for the second population. For the third and the fourth population, study and auxiliary variables are generated from different distributions.

Table 1. Distributions of study and auxiliary variables for each population

Stratum	Correlation Within Stratum	
Population one		
h=1	ρ_{yx_h}=0.5	$N(0,10) = f(y_1) = \frac{1}{10\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{y_1}{10}\right)^2}{2}\right]$
h=2	ρ_{yx_h}= 0.7	$N(0,6) = f(y_2) = \frac{1}{6\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{y_2}{6}\right)^2}{2}\right]$
h=3	ρ_{yx_h}= 0.9	$N(0,5) = f(x_1) = \frac{1}{5\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{x_1}{5}\right)^2}{2}\right]$ $N(0,3) = f(x_2) = \frac{1}{3\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{x_2}{3}\right)^2}{2}\right]$
Population Two		
h=1	ρ_{yx_h}=0.5	$x^2(7) = f(y_1) = \frac{1}{2^{\frac{7}{2}}\,\Gamma\frac{7}{2}} {y_1}^{\frac{7}{2}-1} \exp\left[-\frac{(y_1)}{2}\right]$
h=2	ρ_{yx_h}= 0.7	$x^2(9) = f(y_2) = \frac{1}{2^{\frac{9}{2}}\,\Gamma\frac{9}{2}} {y_2}^{\frac{9}{2}-1} \exp\left[-\frac{(y_2)}{2}\right]$
h=3	ρ_{yx_h}= 0.9	$x^2(2) = f(x_1) = \frac{1}{2^{\frac{2}{2}}\,\Gamma\frac{2}{2}} {x_1}^{\frac{2}{2}-1} \exp\left[-\frac{(x_1)}{2}\right]$ $x^2(6) = f(x_2) = \frac{1}{2^{\frac{6}{2}}\,\Gamma\frac{6}{2}} {x_2}^{\frac{6}{2}-1} \exp\left[-\frac{(x_2)}{2}\right]$
Population Three		
h=1	ρ_{yx_h}=0.5	$N(0,5) = f(y_1) = \frac{1}{5\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{y_1}{5}\right)^2}{2}\right]$
h=2	ρ_{yx_h}= 0.7	$N(0,3) = f(y_2) = \frac{1}{3\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{y_2}{3}\right)^2}{2}\right]$
h=3	ρ_{yx_h}= 0.9	$\Gamma(1.2,1) = f(x_1) = \frac{1}{\Gamma(1.2)} {x_1}^{1.2-1} \exp(-x_1)$ $\boldsymbol{x^2}(9) = f(x_2) = \frac{1}{2^{\frac{9}{2}}\,\Gamma\frac{9}{2}} {x_2}^{\frac{9}{2}-1} \exp\left[-\frac{(x_2)}{2}\right]$
Population Four		
h=1	ρ_{yx_h}=0.5	$\Gamma(1.2,1) = f(y_1) = \frac{1}{\Gamma(1.2)} {y_1}^{1.2-1} \exp(-y_1)$
h=2	ρ_{yx_h}= 0.7	$\boldsymbol{x^2}(9) = f(y_2) = \frac{1}{2^{\frac{9}{2}}\,\Gamma\frac{9}{2}} {y_2}^{\frac{9}{2}-1} \exp\left[-\frac{(y_2)}{2}\right]$
h=3	ρ_{yx_h}= 0.9	$N(0,5) = f(x_1) = \frac{1}{5\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{x_1}{5}\right)^2}{2}\right]$ $N(0,3) = f(x_2) = \frac{1}{3\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{x_2}{3}\right)^2}{2}\right]$

6.3. The Used Software Packages

Three software packages are used in this study; R, GAMS, and Microsoft Excel. These three packages were combined together in order to perform the required simulation study as follows:

R:

It is used to generate (for each population) 1000 datasets; each contains 2 study variables and 2 auxiliary variables. They are generated according to the parameter values mentioned earlier in table 1. The number of observations generated per variable is 10,000 which is the total population size. A 1000 sized sample is proportionally selected from each population under stratified sampling scheme. These calculated values for the means and the variances represent the input for the models under comparison. From each dataset, a proportional stratified random sample is drawn. Then R software calculates, for each population and its selected sample, the means and the variances of the four generated variables within the three strata, and then writes them to an excel file. These excel files represent the input for the GAMS software.

Microsoft Excel:

Excel is used as an intermediate program between GAMS and R in addition to calculation of the Relative Root Mean Square Error (RRMSE %) criteria.

GAMS:

GAMS is used to solve the different proposed GP models and the models presented in the literature using the inputs saved in an excel files (a CSV file). It solves the nonlinear programming problems using CONOPT 3.15L solver (by default). Then, GAMS writes the output of the models into a GDX file. Finally this GDX file can be rewritten in a new excel file in which the RRMSE% criteria is calculated for each considered scenario.

6.4. Simulation Results

In order to assess the performance of the proposed approach for the GCE, the proposed GP model is solved in different scenarios for each generated population. For each population from the four generated populations, the GP model is solved in 12 scenarios, in addition to solving the models presented in the literature. So it can be said that the model is solved in around 48 different scenarios with 1000 iterations in each scenario. The different 12 scenarios can be detailed as follows:

1. A *multivariate* model with two *correlated* study variables when both of the variance and the bias of each calibrated estimator has the same relative importance.
2. A *multivariate* model with two *correlated* study variables when the variance of each calibrated estimator is relatively more important than its unbiasness.
3. A *multivariate* model with two *correlated* study variables when getting an unbiased calibrated estimator is relatively more important than decreasing its variance
4. A *multivariate* model with two *uncorrelated* study variables when both of the variance and the bias of each calibrated estimator has the same relative importance.
5. A *multivariate* model with two *uncorrelated* study variables when the variance of each calibrated estimator is relatively more important than its unbiasness.
6. A *multivariate* model with two *uncorrelated* study variables when getting an unbiased calibrated estimator is relatively more important than decreasing its variance.
7. A *univariate* model focusing on the first study variable Y1 when both of the variance and the bias of each calibrated estimator has the same relative importance.
8. A *univariate* model focusing on the first study variable Y1 when the variance of each calibrated estimator is relatively more important than its unbiasness.
9. A *univariate* model focusing on the first study variable Y1 when getting an unbiased calibrated estimator is relatively more important than decreasing its variance.
10. A *univariate* model focusing on the second study variable Y2 when both of the variance and the bias of each calibrated estimator has the same relative importance.
11. A *univariate* model focusing on the second study variable Y2 when the variance of each calibrated estimator is relatively more important than its unbiasness.
12. A *univariate* model focusing on the second study variable Y2 when getting an unbiased calibrated estimator is relatively more important than decreasing its variance.

Hereafter a comparison between the different considered scenarios using the RRMSE% for the four generated populations is presented.

Table 2. Simulation Results for the First Generated Population.

Number of study variables	Model Design	Relative Importance	RRMSE% ($\bar{y}_1^c$)	RRMSE% ($\bar{y}_2^c$)
Multivariate	Correlated	Case 1:	9906.87	3691.665
		Case 2	9910.60	3677.516
		Case 3	12166.43	7694.158
	Uncorrelated	Case 1	9904.337883	3685.530338
		Case 2	9910.607536	3677.516327
		Case 3	12234.90437	6812.803631
Univariate	Min V($\bar{y}_{1st}^c$)	Case 1	9896.624221	5673.716193
		Case 2	12237.04238	6556.127475
		Case 3	9896.431891	5568.746445
	Min V($\bar{y}_{2st}^c$)	Case 1	12240.72157	3652.183657
		Case 2	12236.24	6523.869686
		Case 3	14914.85521	3650.952088
Some Literature	Tracy [4]		6054.086	3663.296
	Singh [3]		13351.07	7235.36
	Rao [5]		5260.626392	2538.814
	Rao [9]		15158.22745	7158.523107

- ***The First Generated Population:***

With respect to the 1st generated population, in which both of the study and auxiliary variables are generated from normal distribution with different parameters, hereafter there are six sides of comparison using RRMSE% presented in table 2. ***The first side*** represents a comparison between the efficiency of the three multivariate calibrated estimators when having two correlated study variables, according to the priority weights given to the variance and bias goal constraints.

In this side when comparing the efficiency of the three proposed models according to the relative importance of the variance and biasness of the calibrated estimators, it can be noticed that the 1st calibrated estimator $\bar{y}_1^c$ is more efficient in case 1 than the other two cases. In other words, $\bar{y}_1^c$ has a lower RRMSE% in the case where same priorities are given to both of the variance and biasness goal constraints. On the other hand, the 2nd calibrated estimator $\bar{y}_2^c$ has a lower RRMSE% in case 2 when its variance is considered more relatively important than its biasness.

Regarding ***the second side***, it represents also a comparison between the efficiency of the three multivariate calibrated estimators based on the priority weight of each goal constraint. But it focuses on the case when the correlation between the two study variables is ignored. It is observed from table 2 that the same previous manner of RRMSE% criteria is found in the proposed models when the correlation between the two study variables is ignored.

Contrary to the previous two sides, ***the third side*** represents a comparison based on the correlation structure between the two study variables. So this side is considered to assess to what extent the correlation structure between the study variables may affect the estimator's efficiency. According to the simulation results presented in table 2, it can be shown that the 1st calibrated estimator $\bar{y}_1^c$ is more efficient when ignoring the correlation between the two study variables. While the efficiency of the 2nd calibrated estimator $\bar{y}_2^c$ is almost the same in both cases when ignoring the correlation and when considering it.

On the other hand, the fourth and fifth sides are related to the comparison between the proposed models when the variance of each calibrated estimator is minimized separately in a univariate model. The comparison here is carried out according to the priority weights given to the variance and bias goal constraints.

The fourth side is a comparison between the RRMSE% criteria of the calibrated estimators when focusing on minimizing the variance of the 1st calibrated estimator only. It can be noted that both of $\bar{y}_1^c$ and $\bar{y}_2^c$ have a lowest RRMSE% in case 3. So the proposed calibrated estimators are more efficient when a higher priority is given to the biasness goal constraint than variance goal constraint.

Similarly, the ***fifth side*** represents a comparison, based on the priority weights given to each goal constraint, between the three univarite models when focusing on minimizing the variance of the 2nd calibrated estimator. From table 2 it can be noted that the 1st calibrated estimator $\bar{y}_1^c$ performed better in case 2 as its RRMSE% is lower than the other two cases. While the 2nd calibrated estimator $\bar{y}_2^c$ has the same efficiency as its efficiency in the fourth side since the RRMSE% ($\bar{y}_2^c$) is still the lowest in case 3.

Finally, the ***sixth side*** is recommended to compare the most efficient calibrated estimator, among the 12 proposed scenarios, with some literature. For the first calibrated

estimator, it has the lowest RRMSE%, among all the proposed 12 scenarios, in case 3 when focusing on minimizing its variance only. Comparing the most efficient estimator, for the first study variable, by the literature estimators, it is noticed that the proposed calibrated estimator is more efficient than the calibrated estimators of [3] and [9]. But the calibrated estimators of Tracy [4] and Rao [5] performed better than it.

Regarding the second calibrated estimator, it is logically to find that it is the most efficient, among the 12 proposed scenarios, when focusing on minimizing its variance in a univariate model separately. By comparing it with some literature studies, it exceeds in efficiency the calibrated estimators of [4], [3], and [9].

- ***The Second Generated Population:***

In the second generated population, both of the study variables and auxiliary variables are generated from chi-square distribution with different degrees of freedom as shown before in table 1. The simulation results for this population, expressed in the RRMSE% criteria, are given in table 3.

By applying the previously detailed comparison sides using the RRMSE% criteria of the proposed calibrated estimators, it can be said that there is no obvious difference in the RRMSE% between the considered different scenarios. This means that the proposed calibrated estimators have the same efficiency level, in the different scenarios, when generating the data from chi-square distribution.

Through comparing the proposed estimators by the literature, it can be noted that the 1st proposed calibrated estimator performed better than the estimator proposed by [4]. Moreover both of [3] and [9]'s calibrated estimators have almost the same efficiency as the efficiency of the proposed estimators. While [5]'s estimator has a lower RRMSE than the proposed estimators. With respect to the 2nd calibrated estimators, it is notable that the proposed approach for GCE is more efficient than the calibrated estimator of [3], [4] and [9].

Table 3. Simulation Results for the Second Generated Population

Number of study variables	Model Design	Relative Importance	RRMSE% ($\bar{y}_1^{\,c}$)	RRMSE% ($\bar{y}_2^{\,c}$)
Multivariate	Correlated	Case 1	1.562059276	1.103731163
		Case 2	1.562059276	1.103731163
		Case 3	1.562059276	1.103731163
	Uncorrelated	Case 1	1.562059276	1.103731163
		Case 2	1.562059276	1.103731163
		Case 3	1.562059276	1.103731163
Univariate	Min V($\bar{y}_{1_{st}}^{\,c}$)	Case 1	1.562059276	1.103731
		Case 2	1.562059276	1.103731163
		Case 3	1.562059276	1.103731
	Min V($\bar{y}_{2_{st}}^{\,c}$)	Case 1	1.562059276	1.103731163
		Case 2	1.562059276	1.103731163
		Case 3	1.562059276	1.103731163
Some Literature	Tracy [4]		1.618687	1.498896
	Singh [3]		1.555073	1.164961
	Rao [5]		0.967649	0.679893
	Rao [9]		1.543594	1.281325

Table 4. Simulation Results for the Fourth Generated Population.

Number of study variables	Model Design	Relative Importance	RRMSE% ($\bar{y}_1^c$)	RRMSE% ($\bar{y}_2^c$)
Multivariate	Correlated	Case 1	2540.341312	1814.743
		Case 2	2529.676475	1783.42
		Case 3	15238.15114	1649.837069
	Uncorrelated	Case 1	2532.715	1797.906
		Case 2	2529.676	1783.42
		Case 3	15238.15	16498.37
Univariate	Min V($\bar{y}_{1st}^c$)	Case 1	2527.075086	9640.787
		Case 2	15238.15114	16498.37069
		Case 3	2523.677394	9446.543284
	Min V($\bar{y}_{2st}^c$)	Case 1	20452.2	1771.064
		Case 2	15238.15	16498.37
		Case 3	20148.67	1765.515
Some Literature	Tracy [4]		18443	15518.06474
	Singh [3]		17539.22974	16390.0908
	Rao [5]		14134.1	14302.9
	Rao [9]		17593.27	15905.15

- ***The Third Generated Population:***

In the third generated population, the two study variables are generated from normal distribution with different parameters. One of the auxiliary variables is generated from Gamma distribution, and the second from chi-square with the parameters shown in table 1. Table 4 presents the simulation results for this population from which it can be noted that there are different levels for the RRMSE% criteria in the considered different scenarios. In the following paragraphs, a detailed comparison will be presented.

Considering the previously detailed sides of comparisons, in ***the first side*** when comparing the efficiency of the three multivariate models according to the relative importance of the variance and biasness goal constraints, it can be noticed that the 1st calibrated estimator is more efficient in case 2 than the other two cases. Sine $\bar{y}_1^c$ has a lower RRMSE% in the case where the variance of the calibrated estimators is considered more relatively important than its biasness. On the other hand, the 2nd calibrated estimator $\bar{y}_2^c$ has a lower RRMSE% in case 3 when the biasness of the calibrated estimators is considered more relatively important than its variance.

Regarding ***the second side***, in which the efficiency of the three multivariate calibrated estimators is compared ignoring the correlation between the study variables, it is observed from table 4 that both of the first and the second calibrated estimator are more efficient in case 2 than other two cases.

Contrary to the previous two sides, ***the third side*** is considered for assessing which case the estimator is more efficient, either when considering the correlation between the study variables or when ignoring it. It can be shown from table 4 that the 1st calibrated estimator $\bar{y}_1^c$ has the same efficiency level in both cases. While the 2nd calibrated estimator $\bar{y}_2^c$ is more efficient when considering the correlation between the study variables

Since the fourth and the fifth sides focus on the comparison between the proposed univariate models when the variance of each calibrated estimator is minimized separately. The comparison here is carried out according to the priority weights given to the variance and bias goal constraints.

In ***the fourth side***, a comparison between the RRMSE% criteria of the calibrated estimators when focusing on minimizing the variance of the 1st calibrated estimator only is carried out. It can be noted that both of $\bar{y}_1^c$ and $\bar{y}_2^c$ have a lowest RRMSE% in case 3. So the proposed calibrated estimators are more efficient when a higher priority is given to the biasness goal constraint than the variance goal constraint.

Similarly, the ***fifth side*** represents a comparison, based on the priority weights given to each goal constraint, between the three univarite models when focusing on minimizing the variance of the 2nd calibrated estimator. It can be noted from table 4 that the 1st calibrated estimator $\bar{y}_1^c$ performed better in case 2 than the other two cases. While the 2nd calibrated estimator $\bar{y}_2^c$ has the same efficiency as its efficiency in the fourth side since the RRMSE% ($\bar{y}_2^c$) is still the lowest in case 3.

Finally, the ***sixth side*** is recommended to compare the most efficient calibrated estimator, among the 12 proposed scenarios, by the calibrated estimators presented in the literature. For the first calibrated estimator it has the lowest RRMSE% in case 3 when focusing on minimizing its variance only. Compared by the literature estimators, the

first calibrated estimator is more efficient than the estimators of [3], [4], [5] and [9]. So it can be said that the proposed approach for MCE outperformed all the considered literature calibrated estimators in the third generated population.

Regarding the second calibrated estimator, it unexpectedly has the lowest RRMSE%, among the 12 proposed scenarios, when considering a correlated multivariate model and giving the biasness goal a higher priority than the variance goal. By comparing the most efficient 2nd calibrated estimator by some literature, it is found that it exceeds in efficiency all the considered studies of the literature.

- ***The fourth Generated Population :***

Regarding the fourth generated population, it can be considered the inverse case of the third population. Since the two study variables are generated from gamma and chi-square distribution, while the two auxiliary variables follow Normal distribution with different parameters. The simulation results expressed in the RRMSE% criteria are given in table 5.

By applying the previous sides of the comparison depending on the RRMSE% criteria of the proposed calibrated estimators, it can be noted that there is no obvious difference in the RRMSE% between the considered 12 different scenarios. This means that all the proposed calibrated estimators have the same efficiency level for the fourth generated population.

Through comparing the efficiency level of the proposed approach by the literature calibrated estimators, it can be noted that the proposed approach for MCE for the 1st calibrated estimator exceeds in the efficiency both of [4], [5] and [9]'s calibrated estimators. While Singh [3]'s calibrated estimator has almost the same RRMSE% as the proposed estimators. On the other hand the 2nd calibrated estimator has a higher efficiency level than all the considered calibrated estimators in the literature.

7. The Application of the Proposed Approach

This section presents an application of the proposed calibration approach using the sampling survey of the Health and Social Conditions of the Elderly (HSCE) in order to coordinate its results with available high-quality external information to increase their representativeness and usefulness. Since the external information is supposed to be taken from a high quality data source, the 2017 Egypt population census results will be considered the source of the auxiliary information.

Regarding the HSCE sampling survey, it collects data, using the household questionnaire. The household questionnaire is used in many social surveys, such as the Income and Expenditure Survey, the Demographic and Health Survey, and the Health and Social Conditions Survey for the elderly. It collects data on all family members in terms of age, education and work, as well as data on housing conditions and family property. The HSCE sampling survey collects data on the elderly in terms of age, household size, the number of rooms in the house, the age, and finally the region variable that is used to determine the strata. Accordingly two strata will be considered in this application according to the region variable (Urban and Rural).

Table 5. Simulation Results for the Third Generated Population.

Number of study variables	Model Design	Relative Importance	RRMSE% ($\bar{y}_1^c$)	RRMSE% ($\bar{y}_2^c$)
Multivariate	Correlated	Case 1	2.861568	1.374593
		Case 2	2.861568	1.374593
		Case 3	2.861568	1.374593
	Uncorrelated	Case 1	2.861568208	1.374593032
		Case 2	2.861568	1.374593032
		Case 3	2.861568	1.374593032
Univariate	Min V($\bar{y}_{1st}^c$)	Case 1	2.861568208	1.374593
		Case 2	2.861568208	1.374593
		Case 3	2.861568	1.374593
	Min V($\bar{y}_{2st}^c$)	Case 1	2.861568208	1.374593032
		Case 2	2.861568208	1.374593
		Case 3	2.861568208	1.374593032
Some Literature	Tracy [4]		3.882574	2.612947
	Singh [3]		2.860164	1.394817
	Rao [5]		2.9381	1.901396
	Rao [9]		2.86235	1.921602

7.1. The Proposed Study and Auxiliary Variables

The main study variable that will be considered in this application is the individual's age, assuming that we are interested in estimating its mean $\overline{Y}$ in the HSCE survey by incorporating the number of HH members; X1 and the number of rooms per household; X2 as the auxiliary variables from the census results. In other words, the application aims to use the suggested goal programming model for minimizing the variance of the calibrated estimator of the mean of the individual's age. In addition to that, it aims in the consistency of the results of the census data for the variables of the number of HH members and the number of rooms per household, by adjusting the urban and rural weights. Tables 6 and 7 present the main features of the considered variables for the application from the survey and the census.

Table 6. Needed Information from the HSCE survey

Strata (h; 1,2)	n_h	f_h	$s^2{}_{y_h}$	$\overline{x}_{h1}$	$\overline{x}_{h2}$	$s^2{}_{x_2}$
Urban	179	0.0000013	78.261	3.25	3.23	2.69
Rural	120	0.0000022	65.76	3.43	3.39	3.8
Total	n =299					

Table 7. Basic Features of the Population Information

Stratum	N_h	W_h; design weights	Some Population Parameters
Urban 1	40,192,691	0.4	μ_{x1}= 4.04
Rural 2	54,564,390	0.6	μ_{x2}= =3.42
	N =94,757,081	1	$\sigma^2{}_{x_2}$ = 0.577

7.2. The Proposed Model Application

Since the application of the proposed calibration approach will concentrate on the case of having one study variable incorporating two auxiliary variables, the main structure for the MPM's application can be expressed as follows:

Find Ω_h that

$$Min\ Z = \sum_{h=1}^{L} |W_h - \Omega_h|, \quad (7.1)$$

Subject to

$$\sum_{h=1}^{L} \Omega_h \overline{x}_{h1} = \overline{X_1}, \quad (7.2)$$

$$\sum_{h=1}^{L} \Omega_h \overline{x}_{h2} = \overline{X_2}, \quad (7.3)$$

$$\sum_{h=1}^{L} \Omega_h s^2{}_{h2} = S^2{}_2\ , \quad (7.4)$$

$$\sum_{h=1}^{L} \Omega_h = 1\ , \quad (7.5)$$

$$\sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy}^2 \le \zeta\ , \quad (7.6)$$

$$\Omega_h \ge 0,\ h=1, 2 \quad (7.7)$$

where;

W_h & Ω_h is the design and calibrated weights of the urban and rural regions, respectively; h=1, 2.

$\overline{x}_{h1}$ is sample mean for the HH size in h[th] stratum; h=1, 2.

$\overline{X}_1$ is mean for the HH size from the census.

$\overline{x}_{h2}$ is the sample mean for the number of house's rooms in h[th] stratum; h=1, 2.

$\overline{X}_2$ is mean for the number of rooms per household from the census.

$s^2{}_{h2}$ is the sample variance for the number of house's rooms in h[th] stratum;

$S^2{}_2$ is the variance of the number of rooms per household from the census.

s_{hy}^2 is the sample variance of the individuals' age in the h[th] stratum; h=1,2.

$\sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy}^2$ denotes the estimated variance of the calibrated estimator.

ζ is a positive upper bound for the estimated variance of the calibrated estimator.

Regarding the solving method of the previous MP model, the same procedure used previously in the simulation study is also used here . So that the final GP model can be expressed as follows:

Find Ω_h, dn_{1h}, dp_{1h} ; h=1,2. dn_k, dp_k; k=2, 3,..5, that

$$\text{Minimize } Z = \sum_{h=1}^{L}(dn_{1h} + dp_{1h}) + dp2 + \sum_{k=3}^{5}(dn_k + dp_k) \quad (7.8)$$

Subject to;

$$[\Omega_h/W_h] + dn_{1h} - dp_{1h} = 1;\ h=1, 2 \quad (7.9)$$

$$[\sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h}\right) s_{hy}^2]/\zeta + dn_2 - dp_2 = 1; \quad (7.10)$$

$$[\sum_{h=1}^{L} \Omega_h \overline{x}_{h1}]/\ \overline{X_1}\ + dn_3 - dp_3 = 1; \quad (7.11)$$

$$[\sum_{h=1}^{L} \Omega_h \overline{x}_{h2}]\ /\ \overline{X_2}\ + dn_4 - dp_4 = 1; \quad (7.12)$$

$$[\sum_{h=1}^{L} \Omega_h s^2{}_{2h}]\ /S^2{}_2 + dn_5 - dp_5 = 1; \quad (7.13)$$

$$\sum_{h=1}^{L} \Omega_h = 1; \quad (7.14)$$

7.3. The Results

The Application of the Proposed Model is solved using GAMS software and the estimated calibration weights for the urban and rural regions are obtained and given in table (8).

Table 8. The Estimated Calibrated Weights

Stratum; h	Design weights	Calibrated weights
1	0.4	0.424
2	0.6	0.576
Sum	1	1

Depending on the values of the design and calibrated weights given in table 8, the estimated variance of the design estimator is calculated by **0.267233** while the

estimated variance of the calibrated estimator is **0.260413**. The comparison between the design and the calibrated estimator can be demonstrated using the Relative Efficiency measure given by [9] as follows:

$$RE= \frac{\hat{v}(\bar{y}_{st})}{\hat{v}(\hat{\bar{Y}})} \times 100 \qquad (7.15)$$

where $\hat{v}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 (\frac{1-f_h}{n_h}) s_{hy}^2$ is the estimated variance of the unbiased estimator $\bar{y}_{st}$, $\hat{v}(\hat{\bar{Y}})$ is the estimated variance of a calibration estimator, that can be computed through substituting by the calibrated weights instead of the design weights in (7.15). It can be noted from the RE measure that the proposed calibration approach has already improved the variance of the design estimator by 2.6 %.

Further, it is suggested to apply some of the calibrated estimators presented in the literature using the same data in order to compare their performance with the proposed calibrated estimator using the RE. Table 9 presents the variance of the calibration estimator $\hat{v}(\hat{\bar{Y}})$, and its relative efficiency (RE) for some different calibrated estimators.

Table 9. The Calibrated Weights, the Estimated Variance, and the RE for Different Calibrated Estimators

	Calibrated weights		Variance	RE%
The design estimator	Urban	0.4	0.26723	100%
	Rural	0.6		
The proposed estimator	Urban	0.424	**0.260413**	**102.6%**
	Rural	0.576		
Singh 2003	Urban	0.395	0.268797	99.4%
	Rural	0.605		
Rao 2012	Urban	0.41	0.282789	94.5%
	Rural	0.618		
Tracy 2003	Urban	0.872549	0.341769	78.2%
	Rural	0.127451		

From table 9, it can be noted that the application of the calibration estimation reveals a better performance for the proposed model than the considered literature's estimators. As it has the highest RE value among all the considered calibrated estimators.

8. Conclusions

In this paper, using GCE approach is newly suggested to be used for improving the precision of the estimates of population parameters of multi study variables by incorporating multi auxiliary variables. Depending on the correlation structure between the study variables, two different mathematical programming models are proposed and solved using GP approach. A simulation study is conducted by simulating four different populations for representing the symmetric, skewed, and mixed distributions. For each generated population, two study variables and two auxiliary variables as a special case for the GCE approach were generated. The main purpose behind the simulation study is assessing the precision of the proposed calibration approach in addition to determine to what extent the correlation structure between the study variables affect its performance. The results showed that when the study variables are generated from Normal distribution regardless the distribution from which the auxiliary variables are generated, different values for the RRMSE% criteria are given for the proposed 12 scenarios. For example in the first generated population, the proposed approach for GCE has a higher performance than the calibrated estimators of both of [3] and [9]. While in the third population, the results showed that the proposed approach of GCE gave the best performance compared by the estimators proposed by both of [3], [4], [5] and [9].

On the other hand, generating the study variables from a skewed distribution gave similar performance levels in the different considered scenarios which can be noticed in the second and fourth populations. Finally it can be noted that the proposed estimators, in the different considered scenarios for the four generated populations, provide better efficiency compared to some existing calibration methods. So using GP approach practically provides more efficient and flexible calibration estimator than other existing calibration estimation in the stratified sampling design.

REFERENCES

[1] J. C. Deville, C. E. Sarndal, Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, Vol. 87, No.418, 376-382, 1992.

[2] S. Singh, S. Horn, F. Yu, Estimation of variance of the general regression estimator: Higher level calibration approach, Survey Methodology, 24, 41-50, 1998.

[3] S. Singh Advanced Sampling Theory with Applications, Kluwer Academic Publishers, 2003.

[4] D.S. Tracy, S. Singh, R. Arnab. Note on calibration in stratified and double sampling, Survey Methodology, Vol. 29, No. 1, pp. 99-104, 2003.

[5] D. K. Rao, M. G. Khan, S. Khan. Mathematical Programming on Multivariate Calibration Estimation in Stratified Sampling, International Journal of Mathematical, Computational, Physical, electrical and Computer Engineering, Vol. 6, No. 12, pp. 58-62, 2012.

[6] S. Rabee, R. Hamed, R. Kassem, M. Rashwaan A Goal Programming Approach for Multivariate Calibration Weights Estimation in Stratified Random Sampling, Mathematics and Statistics, Vol. 9, No. 3, pp. 326-334, 2021. DOI: 10.13189/ms.2021.090314

[7] W.G. Cochran Sampling Techniques, John Wiley and Sons, New York, 1977.

[8] N., Koyuncu, C. Kadılar. A New Calibration Estimator in

Stratified Double Sampling, Hacettepe Journal of Mathematics and Statistics, Vol. 43, No. 2, pp. 1-9, 2016.

[9] D. K. Rao, M. G. Khan, G. Singh On Calibrated Weights in Stratified Sampling, ANZIAM Journal, vol. 59, pp. 190-204, 2018.

[10] M. J. Schniederjans Goal Programming: Methodology and Applications, Kluwer Academic Publishers, 1995.

[11] J. A. Díaz-García, R. Ramos-Quiroga Optimum Allocation in Multivariate Stratified Random Sampling: A Modified Prékopa's Approach, J Math Model Algor, Vol. 13, pp. 315-330, DOI: 10.1007/s10852-013-9238-4, 2014.

[12] N. Özgül New Calibration Estimator Based on Two Auxiliary Variables in Stratified Sampling. Communications in Statistics - Theory and Methods, 1-12, 2018.