scientific reports

Check for updates

OPEN

A general class of improved population variance estimators under non-sampling errors using calibrated weights in stratified sampling

M. K. Pandey[1]✉, G. N. Singh[1], Tolga Zaman[2], Aned Al Mutairi[3] & Manahil SidAhmed Mustafa[4]

This paper proposes a new calibration estimator for population variance within a stratified two-phase sampling design. It takes into account random non-response and measurement errors, specifically applying this method to estimate the variance in Gas turbine exhaust pressure data. The study integrates additional information from two highly positively correlated auxiliary variables to develop a general class of estimators tailored for the stratified two-phase sampling scheme. The properties of these estimators, in terms of their biases and mean square errors, have been thoroughly examined and extensively analyzed through numerical and simulation studies. Furthermore, the calibrated weights of the strata are derived. The proposed estimators outperform the natural estimator of population variance. Finally, suitable recommendations have been made for survey statisticians intending to apply these findings to real-life problems.

In many practical scenarios, estimating population variance is a crucial task with wide-ranging applications, spanning various domains including finance, healthcare, and weather forecasting. Actuaries and insurance analysts heavily rely on population variance estimation to make well-informed decisions. In the realm of weather forecasting, grasping the variability in temperature, humidity, and other meteorological factors at diverse locations is fundamental for precise predictions. To bolster the precision of estimators in sample surveys, auxiliary variables play a pivotal role. For instance, when estimating crop yields, incorporating data on the area covered by crops can significantly enhance prediction accuracy. Numerous studies, such as[1] did work on the use of auxiliary information in estimating the finite population variance[2], developed a class of estimators using auxiliary information for estimating finite population variance, and[3] introduced a new procedure for variance estimation in simple random sampling using auxiliary information[4]. further improved the estimation of finite population variance using dual supplementary information under stratified random sampling, while[6] explored the more efficient use of auxiliary information in population variance estimation, presenting a new family of estimators.

Moreover, recent research has delved into variance estimation using auxiliary information, with innovative approaches like memory type ratio and product estimators[7,8] gaining attention. These endeavors aim to enhance the accuracy and reliability of population variance estimation in diverse sampling designs.

However, sample surveys often encounter practical challenges that result in non-response or missing data. These challenges encompass non-contact, refusal to cooperate, and various other reasons. When a substantial amount of data goes missing, it casts doubt on the reliability of ensuing statistical results. Diverse types of missing data patterns, such as missing at random (MAR) and missing completely at random (MCAR), can be observed. Particularly noteworthy is the MAR pattern, characterized by the probability of missingness being independent of the unobserved data's value.

In the presence of random non-response or measurement errors, various researchers have addressed the need for robust estimators[9]. introduced a class of estimators using auxiliary information for estimating finite

[1]Department of Mathematics and Computing, Indian Institute of Technology (Indian School of Mines), Dhanbad 826004, India. [2]Faculty of Health Sciences, Gumushane University, Gumushane, Turkey. [3]Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. [4]Department of Statistics, Faculty of Science, University of Tabuk, Tabuk, Saudi Arabia. ✉email: maheshbabu3797@gmail.com

population variance in the presence of measurement errors, while[10] developed classes of factor-type estimators in the presence of measurement error[11]. focused on the estimation of the population coefficient of variation in the presence of measurement errors, and[12] worked on estimating the population mean in the presence of measurement error and non-response under stratified random sampling[13]. contributed to the estimation of the finite population distribution function with the dual use of auxiliary information under non-response, and[14] introduced a generalized class of estimators for sensitive variables in the presence of measurement error and non-response[15]. explored the estimation of finite population mean using dual auxiliary variables for non-response using simple random sampling, while[16] and Bhushan (2023) proposed classes of robust estimators to handle correlated measurement errors and new logarithmic type imputation techniques in presence of measurement errors within the survey sampling literature. These errors may stem from flawed measuring instruments, shortcomings in survey methodology, vague questionnaires, or imprecise measurements.

The calibration approach, pioneered by[18], has garnered prominence in statistical practice. Its objective is to devise unbiased estimation procedures with minimal dispersion, leveraging auxiliary variables. Subsequent researchers, exemplified by[19] and[20], have fine-tuned and extended calibration estimation procedures, striving to minimize the divergence between initial and final weights while adhering to calibration equations and constraints.

Recent advances in calibration techniques, as demonstrated by[21], have focused on a class of calibration estimators under stratified random sampling in the presence of various kinds of non-sampling errors[22]. Explored calibration estimation for ratio estimators in stratified sampling for proportion allocation, and[23] further advanced the finite population distribution function estimation with the dual use of auxiliary information under simple and stratified random sampling[5]. investigated the use of dual ancillary variables to estimate the population mean under stratified random sampling, while[24] worked on modified estimators of the finite population distribution function based on the dual use of auxiliary information under stratified random sampling. These techniques have streamlined the optimization of stratum weights in stratified random sampling, ultimately refining estimates, particularly when closely related auxiliary variables are integrated.

To underscore the practical significance of this research, let's consider real-life examples:

1. In healthcare research, when conducting patient surveys to evaluate the effectiveness of medical treatments, not all patients may respond, and measurement errors can occur due to self-reporting. Accurate population variance estimation in such cases is crucial to making informed decisions about treatment strategies.
2. In market research, understanding consumer preferences through surveys is essential for product development and marketing strategies. Non-response from certain demographic groups or errors in survey responses can distort the estimation of market variances, impacting business decisions.
3. In educational assessments, when evaluating the performance of schools or educational programs, student participation may vary, and measurement errors can affect the assessment outcomes. Reliable population variance estimation is vital for making informed policy decisions and improving education quality.
4. By addressing these issues across diverse fields, this innovative framework aims to provide a reliable approach for accurately estimating population variances, thereby enhancing decision-making processes in real-life scenarios. Additionally, the proposed estimation strategy may be applied to estimate the variance in Gas turbine exhaust pressure, as illustrated using real data in a subsequent section of the manuscript.

Survey sampling necessitates addressing uncertainty and imprecision. Neutrosophic statistics, championed by[25] in 'Neutrosophy: Neutrosophic Probability, Set, and Logic: Analytic Synthesis & Synthetic Analysis,' extend classical statistics for indeterminate data. Aslam's contributions include 'A New Sampling Plan using Neutrosophic Process Loss Consideration'[26] and 'Neutrosophic Analysis of Variance: Application to University Students'[27], among others, illustrating its application in handling vague and imprecise observations in populations or samples.

Motivated by the aforementioned discussions, the present work proposes a wide class of estimators of population variance in two-phase sampling for the stratified population in the presence of random non-response and measurement errors in sample data. The stratum weights have been optimized using calibration procedures, which enables us to get more accurate estimates of the population variance. The performances of the suggested class of estimators have been deeply examined through empirical and simulation studies.

Sample structure

Consider a finite population of size N divided into L non-overlapping strata, each containing N_k(k=1,2,..., L) units. Let Y, X, and Z be the study variable, first and second auxiliary variables, respectively. Let y_{ki}, x_{ki}, and z_{ki} be the ith values of y, x, and z for the k-th (k = 1, 2,..., L) stratum. To estimate the population variance of the study variable Y, It is assumed that the information on the second auxiliary Z is readily available for all the population units. Hence its population variance is known. However, information on the first auxiliary variable is not available for all the units of the population. It is also assumed that the random non-response is observed in the sample data on the study and first auxiliary variables Y and X, respectively. In the first phase, a sample, say S_{n_k} of size n_k (k=1,2..., L), is drawn from kth strata using simple random sampling without replacement and observed for the variables y and x. Let in the first phase sample of size n_k, $n_{k-r_{1k}}$ respond and random non-response observed on the r_{1k} units. Again in the second phase, from the $n_{k-r_{1k}}$ respondent units, another simple random sample without replacement, say S_{m_k}, of size m_k, is chosen from which $m_{k-r_{2k}}$ units respond and r_{2k} units do not respond.

Notations

From now on, we will use the following notations:

σ_Y^2: The population variance of Y, i.e, the characteristics under study

$S^2_{Y_{N_k}}=\frac{1}{N_k-1}\sum_{j=1}^{N_k}(Y_{kj}-\bar{Y}_{N_k})^2$: The population mean squares of the kth stratum of the study variable Y.

$S^2_{X_{N_k}}=\frac{1}{N_k-1}\sum_{j=1}^{N_k}(X_{kj}-\bar{X}_{N_k})^2, S^2_{Z_{N_k}}=\frac{1}{N_k-1}\sum_{j=1}^{N_k}(Z_{kj}-\bar{Z}_{N_k})^2$: The population mean squares of the kth stratum for the auxiliary variables X and Z, respectively.

$s^{*^2}_{x_{n_k}}=\frac{1}{n_k-r_{1k}-1}\sum_{j=1}^{n_k-r_{1k}}(x_{kj}-\bar{x}_{n_k-r_{1k}})^2$:Depending on the responding part of sample S_{n_k}, the sample mean square of auxiliary variable X for the kth stratum.

$s^{*^2}_{x_{m_k}}=\frac{1}{m_k-r_{2k}-1}\sum_{j=1}^{m_k-r_{2k}}(x_{kj}-\bar{x}_{m_k-r_{2k}})^2$: Depending on the responding part of sample S_{m_k}, the sample mean square of auxiliary variable X for the kth stratum.

$s^{*^2}_{y_{n_k}}=\frac{1}{n_k-r_{1k}-1}\sum_{j=1}^{n_k-r_{1k}}(y_{kj}-\bar{y}_{n_k-r_{1k}})^2$: Depending on the responding part of sample S_{n_k}, the sample mean square of study variable Y for the kth stratum.

$s^{*^2}_{y_{m_k}}=\frac{1}{m_k-r_{2k}-1}\sum_{j=1}^{m_k-r_{2k}}(y_{kj}-\bar{y}_{m_k-r_{2k}})^2$: Depending on the responding part of sample S_{m_k}, the sample mean square of study variable Y for the kth stratum.

$s^{*^2}_{z_{n_k}}=\frac{1}{n_k-r_{1k}-1}\sum_{j=1}^{n_k-r_{1k}}(z_{kj}-\bar{z}_{n_k-r_{1k}})^2$: Depending on the responding part of sample S_{n_k}, the sample mean square of auxiliary variable Z for the kth stratum.

$s^{*^2}_{z_{m_k}}=\frac{1}{m_k-r_{2k}-1}\sum_{j=1}^{m_k-r_{2k}}(z_{kj}-\bar{z}_{m_k-r_{2k}})^2$: Depending on the responding part of sample S_{m_k}, the sample mean square of auxiliary variable Z for the kth stratum.

$W_k=\frac{N_k}{N}$: The original weight of the kth stratum, k= 1, 2,....,L

W^*_k: The weight obtained by calibration of the kth stratum, k= 1, 2,....,L

Q_k: The independent weight of the kth stratum, k= 1, 2,....,L

Non-response probability model

The kth stratum is considered based on the random non-response model proposed by Singh and Joarder[28]. In the first phase, a sample of size n_k taken from the population, $n_k - r_{1k}$ units responded, while random non-response was observed on the remaining r_{1k} units, where r_{1k} may have any value from the set $\{0, 1, 2, ..., (n_k - 2)\}$. Again, in the second phase, from the $n_k - r_{1k}$ respondent units, $m_k - r_{2k}$ units responded, and r_{2k} do not respond, where r_{2k} falls within the range $\{0, 1, 2, ..., (m_k - 2)\}$. It is assumed that $r_{jk} \geq 0,\ j = 1, 2$ and $r_{1k} \leq (n_k - 2)$, $r_{2k} \leq (m_k - 2)$. Non-response may have possible values of $(n_k - 2)$ and $(m_k - 2)$ in the samples S_{n_k} and S_{m_k}, respectively. These probabilities will be referred to as p_1 and p_2. The total number of ways to obtain $r_{jk}\ (j = 1, 2)$ non-responses is $\binom{n_k-2}{r_{1k}}$ and $\binom{m_k-2}{r_{2k}}$. Then, the discrete random variables r_{1k} and r_{2k} have the corresponding probability distributions shown below:

$$P(r_{1k})=\frac{n_k-r_{1k}}{n_kq_1+2p_1}\binom{n_k-2}{r_{1k}}p_1^{r_{1k}}q_1^{n_k-r_{1k}-2}\ ;\ r_{1k}=0,1,2,...,n_k-2 \text{ and}$$

$$P(r_{2k})=\frac{m_k-r_{2k}}{n_kq_2+2p_2}\binom{m_k-2}{r_{2k}}p_2^{r_{2k}}q_2^{m_k-r_{2k}-2}\ ;\ r_{2k}=0,1,2,...,m_k-2$$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Suggested estimator

A wide class of estimators that may be used to estimate the population variance are proposed as follows, assuming the impact of random non-response on both the study variable Y and the first auxiliary variable X.

$$T=\sum_{k=1}^{L}W_k^{*2}T_k \tag{1}$$

where

$$T_k=f(s^{*2}_{y_{m_k}},s^{*2}_{x_{m_k}},h(s^{*2}_{x_{n_k}},s^2_{z_{n_k}})),k=1,2,...,L \tag{2}$$

In this case, $h(s^{*2}_{x_{n_k}},s^{*2}_{z_{n_k}})$ is a class of estimators of S^2_X based on information on $s^{*2}_{x_{n_k}}$ and $s^2_{z_{n_k}}$ such that $h(S^2_X,S^2_Z)=S^2_X$.

As we proceed, we will examine the composite class of estimators applicable to individual strata in two-phase sampling.

$$T_k=f(s^{*2}_{y_{m_k}},s^{*2}_{x_{m_k}},h(s^{*2}_{x_{n_k}},s^2_{z_{n_k}}))=g(s^{*2}_{y_{m_k}},s^{*2}_{x_{m_k}},s^{*2}_{x_{n_k}},s^2_{z_{n_k}}) \tag{3}$$

such that $g(S^2_{Y_k},S^2_{X_k},S^2_{X_k},S^2_{Z_k})=S^2_{Y_k}$

We assume that $g(s^{*2}_{y_{m_k}},s^{*2}_{x_{m_k}},s^{*2}_{x_{n_k}},s^2_{z_{n_k}})$ meets the regularity conditions listed below:

- Regardless of the sample chosen, the function $g(s^{*2}_{y_{m_k}},s^{*2}_{x_{m_k}},s^{*2}_{x_{n_k}},s^2_{z_{n_k}})$ takes on values within a closed convex subspace of the four-dimensional real space R^4 that includes the point $(S^2_{Y_k},S^2_{X_k},S^2_{X_k},S^2_{Z_k})$.
- In R^4, the function $g(s^{*2}_{y_{m_k}},s^{*2}_{x_{m_k}},s^{*2}_{x_{n_k}},s^2_{z_{n_k}})$ is continuous and bounded.
- The partial derivatives of $g(s^{*2}_{y_{m_k}},s^{*2}_{x_{m_k}},s^{*2}_{x_{n_k}},s^2_{z_{n_k}})$ of the first, second, and third orders exist and are continuous and bounded in R^4.

The class of estimators T_k is extensive, as any parametric function $g(s^{*2}_{y_{m_k}}, s^{*2}_{x_{m_k}}, s^{*2}_{x_{n_k}}, s^{2}_{z_{n_k}})$ that meets the stated regularity conditions, and has $g(S^2_{Y_k}, S^2_{X_k}, S^2_{X_k}, S^2_{Z_k}) = S^2_{Y_k}$, may generate estimators for the population mean square of each stratum. Several examples of this class of estimators are:

$$T_{1k} = \frac{s^{*2}_{y_{m_k}}}{s^{*2}_{x_{n_k}}}\left[\frac{s^{*2}_{x_{n_k}}}{s^{*2}_{z_{n_k}}}\right]s^2_{Z_k}, \quad T_{2k} = \frac{s^{*2}_{y_{m_k}}}{s^{*2}_{x_{m_k}}}\left[\frac{s^{*2}_{x_{n_k}} s^{*2}_{z_{n_k}}}{s^2_{Z_k}}\right], \quad T_{3k} = s^{*2}_{y_{m_k}} + b_1\left[\left(s^{*2}_{x_{n_k}} + b_2(s^2_{Z_k} - s^{*2}_{z_{n_k}}\right) - s^{*2}_{x_{m_k}}\right]$$

$$T_{4k} = \frac{s^{*2}_{y_{m_k}}}{s^{*2}_{y_{n_k}}}\left[s^{*2}_{x_{n_k}} + b_3(s^2_{Z_k} - s^{*2}_{z_{n_k}}\right] \text{where } b_1, b_2 \text{ and } b_3 \text{ are the true scalars.}$$

$$T_{5k} = s^{*2}_{y_{m_k}}\, exp\left(\frac{s^{*2}_{x_{n_k}} - s^{*2}_{x_{m_k}}\frac{s^2_{Z_k}}{s^{*2}_{z_{n_k}}}}{s^{*2}_{x_{n_k}} + s^{*2}_{x_{m_k}}\frac{s^2_{Z_k}}{s^{*2}_{z_{n_k}}}}\right) \quad \forall k = 1, 2, \ldots L$$

Calibration techniques have been proposed to acquire the optimum strata weights

The new calibration estimator of the population variance under stratified sampling is provided by

$$T = \sum_{k=1}^{L} W^{*2}_k T_k$$

where $T_k = f(s^{*2}_{y_{m_k}}, s^{*2}_{x_{m_k}}, h(s^{*2}_{x_{n_k}}, s^2_{z_{n_k}})), k = 1, 2, ..., L$ and we obtain the calibrated strata weights W^*_k, where k$\in \{1, 2, ..., L\}$.

Based on the following calibration requirements, the distance function (chi-square type) $\sum_{k=1}^{L} \frac{(W^*_k - W_k)^2}{Q_k W_k}$ is minimized:

1. $\sum_{k=1}^{L} W^*_k = 1$
2. $\sum_{k=1}^{L} W^*_k c_{z_k} = C_Z$
3. $\sum_{k=1}^{L} W^*_k c_{x_{m_k - r_{2k}}} = \sum_{k=1}^{L} W_k c_{x_{n_k - r_{1k}}}$

where, $c_{z_k} = \frac{s_{z_k}}{z_k}$, $C_Z = \frac{S_Z}{\bar{Z}}$, $c_{x_{n_k - r_{1k}}} = \frac{s_{x_{n_k - r_{1k}}}}{\bar{x}_{n_k - r_{1k}}}$ and $c_{x_{m_k - r_{2k}}} = \frac{s_{x_{m_k - r_{2k}}}}{\bar{x}_{m_k - r_{2k}}}$.

It is important to note that $Q_k > 0$ are appropriately determined weights that will determine the estimator form.

In Appendix A, detailed derivations have been given.

Bias and mean square error of the suggested estimator

We utilize the transformations provided below while taking into account large sample assumptions to analyze the properties of estimator T:

$$s^{*2}_{y_{m_k}} = S^2_{Y_k}(1 + \varepsilon_{0k}), \quad s^{*2}_{x_{m_k}} = S^2_{X_k}(1 + \varepsilon_{1k}), \quad s^{*2}_{z_{n_k}} = S^2_{Z_k}(1 + \varepsilon_{2k}), \quad s^{*2}_{x_{n_k}} = S^2_{X_k}(1 + \varepsilon_{3k})$$

such that $|\epsilon_{ik}| \leq 1$, ∀i= 0, 1, 2, 3 and $E(\epsilon_{ik}) = 0$.

According to calculations, the Bias(T) and the MSE(T) of the suggested estimator T, which are accurate to the first order of approximation, are as follows:

$$Bias(T) = \frac{1}{2}\sum_{k=1}^{L} W^{*2}_k \begin{bmatrix} S^4_{X_k}\left(f_{1k}d_{22k} + d_{33k}f_{3k} + 2d_{23k}f_{3k}\right)C^2_{1k} + 2S^2_{Y_k}S^2_{X_k}\rho_{01k}\left(d_{12k}f_{1k} + d_{13k}f_{3k}\right) \\ +2S^2_{X_k}S^2_{Z_k}\rho_{12k}\left(d_{24k}f_{2k} + d_{34k}f_{2k}\right) + 2S^2_{Y_k}S^2_{Z_k}d_{14k}f_{2k}\rho_{02k} + S^4_{Z_k}d_{44}f_{2k}C^2_{2k} \end{bmatrix} \quad (4)$$

and

$$MSE(T) = \sum_{k=1}^{L} W^{*4}_k \left[S^4_{Y_k}f_{1k}C^2_{0k} + d^2_{2k}S^4_{X_k}C^2_{1k}f_{4k} + d^2_{4k}S^4_{Z_k}f_{2k}C^2_{2k} + 2d_{4k}\rho_{02k}f_{2k}S^2_{Y_k}S^2_{Z_k} + 2S^2_{Y_k}S^2_{X_k}d_{2k}\rho_{01k}f_{4k}\right] \quad (5)$$

where

$C^2_{0k} = \lambda_{400k} - 1$, $C^2_{1k} = \lambda_{040k} - 1$, $C^2_{2k} = \lambda_{004k} - 1$

$\rho_{01k} = \lambda_{220k} - 1$, $\rho_{02k} = \lambda_{202k} - 1$, $\rho_{12k} = \lambda_{022k} - 1$

$f_{1k} = \left(\frac{1}{m_k q_2 + p_2} - \frac{1}{N_k}\right)$, $f_{2k} = \left(\frac{1}{n_k} - \frac{1}{N_k}\right)$, $f_{3k} = \left(\frac{1}{n_k q_1 + p_1} - \frac{1}{N_k}\right)$

and

$$\lambda_{\alpha\beta\gamma k} = \frac{\mu_{\alpha\beta\gamma k}}{\sqrt{\mu^{\alpha}_{200k}\mu^{\beta}_{020k}\mu^{\gamma}_{002k}}}, \quad \mu_{\alpha\beta\gamma k} = \frac{1}{N_k}\sum_{j=1}^{N_k}(Y_{kj} - \bar{Y}_k)^{\alpha}(X_{kj} - \bar{X}_k)^{\beta}(Z_{kj} - \bar{Z}_k)^{\gamma}$$

Appendix B has detailed derivations.

The suggested estimator's minimum mean square error under optimal condition.

We note from Eq. (5) that the derivatives d_{2k} and d_{4k} have an impact on the MSE of the estimator T. So, in order to acquire the derivatives' optimal values, we minimize the MSE concerning them as follows:

$$d_{2k_{opt}} = -\frac{\rho_{01k}}{C_{1k}^2}\frac{S_{Y_k}^2}{S_{X_k}^2} \tag{6}$$

and

$$d_{4k_{opt}} = -\frac{\rho_{02k}}{C_{2k}^2}\frac{S_{Y_k}^2}{S_{Z_k}^2} \tag{7}$$

We may obtain the minimum mean square error (Min. MSE)) of the estimator T by substituting the optimal values of $d_{2k_{opt}}$ and $d_{4k_{opt}}$ from Eqs. (6) and (7), respectively, in Eq. (5) as follows:

$$Min.MSE(T) = \sum_{k=1}^{L} W_k^{*4} S_{Y_k}^4 \left[f_{1k} C_{0k}^2 - \frac{\rho_{01k}^2}{C_{1k}^2} f_{4k} - \frac{\rho_{02k}^2}{C_{2k}^2} f_{2k} \right] \tag{8}$$

Effect of measurement error

Y and X actual and observed values are denoted by y_{kj_a}, x_{kj_a}, and y_{kj_o}, x_{kj_o}, while u_{kj}, and v_{kj} denote the corresponding measurement errors. Then $x_{kj_a} = x_{kj_o} + v_{kj}$ and $y_{kj_a} = y_{kj_o} + u_{kj}$, resulting in $V(y_{kj_a}) = V(y_{kj_o}) + V(u_{kj})$, with zero covariance term because the errors are independent.

This implies $s_{y_{ka}}^2 = s_{y_{ko}}^2 + s_{u_k}^2$, so that $MSE(s_{y_{ka}}^2) = MSE(s_{y_{ko}}^2) + MSE(s_{u_k}^2)$.

$$\begin{aligned} \therefore Min.MSE(T) &= \sum_{k=1}^{L} MSE(s_{y_{ka}}^2) \\ &= \sum_{k=1}^{L} \left[MSE(s_{y_{ko}}^2) + M(s_{u_k}^2) \right] \\ &= \sum_{k=1}^{L} MSE(s_{y_{ko}}^2) + \sum_{k=1}^{L} MSE(s_{u_k}^2) \end{aligned}$$

The expression for Min.MSE was determined as follows: measurement errors occurred only on the study variable Y and the primary auxiliary variable X, not on the secondary auxiliary variable Z.

$$Min.MSE(T) = \sum_{k=1}^{L} W_k^{*4} S_{Y_k}^4 \left[f_{1k} C_{0k}^2 - \frac{\rho_{01k}^2}{C_{1k}^2} f_{4k} - \frac{\rho_{02k}^2}{C_{2k}^2} f_{2k} \right] + \sum_{k=1}^{L} W_k^{*4} S_{u_k}^4 f_{1k} C_{0k}^{'2} \tag{9}$$

where

$$C_{0k}^{'2} = \lambda_{40k}^{'} - 1, \quad \lambda_{40k}^{'} = \frac{\mu_{40k}^{'}}{\sqrt{\mu_{20k}^{'2}}} \quad and \quad \mu_{abk}^{'} = \frac{1}{N_k} \sum_{j=1}^{N_k} (u_{kj} - \bar{u}_k)^a (v_{kj} - \bar{v}_k)^b$$

Numerical study

An estimator's performance must first be evaluated in terms of its characteristics before it may be used in practical scenarios. Therefore, an empirical investigation has been conducted in this part using both real and simulated data for the suggested estimator.

We are comparing the suggested estimator T and the contemporary estimator τ to see how well they perform in random non-response. The estimator τ is defined as follows:

$\tau = \sum_{k=1}^{L} W_k^{*2} s_{y_{m_k}}^{*2}$

Additionally, we are comparing these estimators with the standard estimator since it is the only available option when dealing with non-response and measurement errors.

The following are the expressions for its MSE, with and without measurement errors, respectively:

$$MSE(\tau) = \sum_{k=1}^{L} W_k^{*4} S_{Y_{N_k}}^4 \left(\frac{1}{m_k q_2 + 2p_2} - \frac{1}{N_k} \right) (\lambda_{400k} - 1) \tag{10}$$

and

$$MSE(\tau) = \sum_{k=1}^{L} W_k^{*4} S_{Y_{N_k}}^4 \left(\frac{1}{m_k q_2 + 2p_2} - \frac{1}{N_k} \right) (\lambda_{400k} - 1) + \sum_{k=1}^{L} W_k^{*4} S_{uk}^4 f_{1k} C_{0k}^{'2} \tag{11}$$

The Percentage Relative Efficiency (PRE) of the proposed estimator T concerning the estimator τ is given by

$PRE = \frac{MSE(\tau)}{Min.MSE(T)} * 100$

Where Eqs. (8)–(11) give the corresponding equations for Min MSE(T) and MSE(τ), without or with measurement errors, respectively.

The following Q_k values have been taken into consideration:
Case A: Q_k=1.0
Case B: Q_k=$\frac{1}{W_k}$
Case C: Q_k=Z_k
Case D: Q_k=$S^2_{Z_{N_k}}$

The calibrated stratum weights and PREs, resulting from both the presence and absence of non-response, are displayed in the tables below, for both simulated and real data.

Study based on simulated data

We conducted a simulation relevant to our theoretical findings using the statistical computing software R. To achieve our objectives, we used the *MASS* package's function *mvrnorm* to generate data from poisson distributions with given parameters and a given correlation coefficient for the study and the auxiliary variables. To generate data from other acceptable distributions, use the function *genCorGen* included in the package *simstudy*. The measurement errors were generated using a univariate standard normal distribution with the function *rnorm*. Table 1 shows the population parameters for the generated data.

The resulting calibrated stratum weights and PREs in presence of non-response and in absence of non-response are shown in Tables 2, 3 and 4, respectively.

Study based on real data

The information in this section demonstrates the practical application of the proposed class of estimators. The dataset utilized is accessible within the UCI machine learning repository, titled "Gas Turbine CO and NOx Emission Data Set." This dataset comprises 36,733 instances featuring 11 sensor measurements from a gas turbine situated in the northwestern region of Turkey, aggregated over an hour using average or sum calculation methods for the analysis of CO and NOx (NO + NO2) flue gas emissions. To conduct the analysis mentioned above, the specific file utilized is $gt_{2011}.csv$.

Parameters	Real data				Simulated data			
	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 1	Stratum 2	Stratum 3	Stratum 4
N_k	2500	2600	2311	8000	10000	15000	5000	8000
n_k	875	910	809	800	3000	3000	2000	800
m_k	625	650	578	500	2000	2000	1200	500
$n_{k-r_{1k}}$	750	780	693	750	2500	2500	1600	750
$m_{k-r_{2k}}$	500	520	462	450	1600	1600	900	450
ρ_{xyk}	0.90	0.89	0.97	0.6	0.9	0.8	0.8	0.6
ρ_{xzk}	0.83	0.92	0.93	0.6	0.9	0.8	0.8	0.6
ρ_{yzk}	0.75	0.83	0.92	0.6	0.9	0.8	0.8	0.6

Table 1. Population parameters..

Case	Stratum	Q_k	W_k	W_k^*
A	1	1	0.2631579	0.44807455
	2	1	0.3947368	0.36962453
	3	1	0.1315789	0.13017106
	4	1	0.2105263	0.05212985
B	1	3.800	0.2631579	0.43883647
	2	2.533	0.3947368	0.38908713
	3	7.600	0.1315789	0.13022077
	4	4.750	0.2105263	0.04185563
C	1	24.990	0.2631579	0.44804409
	2	25.009	0.3947368	0.36968872
	3	25.060	0.1315789	0.13017123
	4	24.970	0.2105263	0.05209596
D	1	25.370	0.2631579	0.44812328
	2	24.740	0.3947368	0.36952187
	3	24.860	0.1315789	0.13017080
	4	25.300	0.2105263	0.05218405

Table 2. Calibrated strata weights for simulated data..

p_1	p_2	In the absence of measurement error				In the presence of measurement error			
		Case A	Case B	Case C	Case D	Case A	Case B	Case C	Case D
0.05	0.05	123.0381	122.6552	123.0369	123.0401	122.9925	122.6105	122.9913	122.9945
0.05	0.10	123.1304	122.7438	123.1291	123.1323	123.0845	122.6989	123.0833	123.0865
0.05	0.15	123.2208	122.8308	123.2195	123.2227	123.1747	122.7857	123.1735	123.1767
0.05	0.20	123.3094	122.9161	123.3081	123.3114	123.2631	122.8707	123.2619	123.2651
0.10	0.05	121.7219	121.3717	121.7208	121.7237	121.6794	121.3300	121.6783	121.6811
0.10	0.10	121.8941	121.5380	121.8929	121.8959	121.8511	121.4959	121.8500	121.8529
0.10	0.15	122.0630	121.7013	122.0619	122.0649	122.0197	121.6588	122.0185	122.0215
0.10	0.20	122.2289	121.8617	122.2277	122.2308	122.1852	121.8189	122.1840	122.1870
0.15	0.05	120.2839	119.9687	120.2828	120.2855	120.2446	119.9302	120.2436	120.2462
0.15	0.10	120.5415	120.2182	120.5404	120.5431	120.5016	120.1791	120.5006	120.5033
0.15	0.15	120.7947	120.4634	120.7936	120.7963	120.7542	120.4238	120.7532	120.7559
0.15	0.20	121.0436	120.7046	121.0425	121.0453	121.0026	120.6644	121.0015	121.0043
0.20	0.05	118.7063	118.4288	118.7054	118.7077	118.6706	118.3937	118.6697	118.6720
0.20	0.10	119.0554	118.7673	119.0544	119.0568	119.0189	118.7314	119.0179	119.0203
0.20	0.15	119.3990	119.1007	119.3981	119.4005	119.3618	119.0641	119.3608	119.3633
0.20	0.20	119.7374	119.4290	119.7364	119.7389	119.6993	119.3917	119.6984	119.7009

Table 3. PRE of T w.r.t. τ for simulated poisson data.

Stratum	PRE (In the absence of measurement error)	PRE (In the presence of measurement error)
Case I	124.2297	124.1813
Case II	123.8176	123.7701
Case III	124.2284	124.18
Case IV	124.2318	124.1834

Table 4. In the absence of non-response, PRE is observed from simulated data when $p_1 = p_2 = 0$..

We employed the subsequent set of primary and auxiliary variables in this study:
Y: Gas turbine exhaust pressure (GTEP)
X: Air filter difference pressure (AFDP)
Z: Turbine inlet temperature (TIT)
The stratification is organized based on the Ambient temperature (AT) in the following manner:
Stratum 1: from 2.1163-12.707 C
Stratum 2: from 12.708-21.759 C
Stratum 3: from 21.760-34.532 C

In real-world circumstances, the goal is to estimate the variance as precisely as possible. However, complete data is typically not always available. Therefore, we consider the case where some data on the study variable is unavailable. The statistical characteristics of the population are detailed in Table 1, while the calibrated weights for the strata are listed in Table 5. The PRE (Precentage Relative Efficiency) for both the non-response and absence of non-response cases is presented in Tables 6 and 7, respectively.

Discussion

After conducting a detailed numerical study, we have identified the following key points:

1. The strata weights produced by the calibration procedures exhibit slight discrepancies from the actual ones, as evident in Tables 2 and 5. Nevertheless, our findings indicate that the calibration technique effectively enhances the stratum weights, resulting in more accurate estimates.
2. Table 3 reveals a consistent pattern: when p_1, $p_2 \in (0.05, 0.1)$, the suggested estimator consistently outperforms the existing estimator, regardless of the presence or absence of measurement errors. This observation is further supported by the real data presented in Table 6.
3. Further analysis of Tables 3 and 6 reveals that an increase in the value of p_2, while keeping p_1 constant, results in a higher PRE. This observation is a significant outcome of our research. Additionally, when p_2 remains fixed and p_1 increases, the PRE decreases, aligning with our expectations.
4. Tables 4 and 7 demonstrate that the proposed estimator yields a higher Percentage Relative Efficiency (PRE) than the conventional estimator in the absence of non-response also, underscoring the effectiveness of our method, even without non-response.
5. It is noteworthy that as the correlation coefficient's value increases, the PRE also increases. Conversely, a decrease in the correlation coefficient leads to a decrease in PRE.

Case	Stratum	Q_k	W_k	W_k^*
A	1	1	0.2061856	0.3209746
	2	1	0.6185567	0.5243500
	3	1	0.1752577	0.1546754
B	1	4.85	0.2061856	0.2566935
	2	1.6167	0.6185567	0.5356455
	3	5.7059	0.1752577	0.2076610
C	1	0.00092	0.2061856	0.3258734
	2	0.00092	0.6185567	0.5187164
	3	0.0009196857	0.1752577	0.1554102
D	1	0.0045425256	0.2061856	0.1816363
	2	0.0031285169	0.6185567	0.6199598
	3	0.0046467310	0.1752577	0.1984039

Table 5. Calibrated strata weights for real data.

		In the absence of measurement error				In the presence of measurement error			
p_1	p_2	Case A	Case B	Case C	Case D	Case A	Case B	Case C	Case D
0.05	0.05	264.5064	280.9632	261.8853	293.3473	262.4120	278.5422	259.8380	290.6966
0.05	0.10	268.6831	284.3495	266.1682	296.1447	266.5021	281.8539	264.0328	293.430
0.05	0.15	272.8552	287.7074	270.4512	298.9049	270.5860	285.1367	268.2260	296.1271
0.05	0.20	277.0228	291.0372	274.7343	301.6286	274.6639	288.3911	272.4175	298.7875
0.10	0.05	241.0445	254.7055	238.8611	264.8419	239.4063	252.8268	237.2576	262.7989
0.10	0.10	246.0068	259.1428	243.8921	268.9041	244.2764	257.1770	242.1959	266.7791
0.10	0.15	251.0106	263.5895	248.9701	272.9584	249.1849	261.5345	247.1779	270.7499
0.10	0.20	256.0565	268.0456	254.0956	277.0046	254.1324	265.8995	252.2041	274.7114
0.15	0.05	219.3091	230.6233	217.4957	238.9029	218.0470	229.1854	216.2590	237.3481
0.15	0.10	224.8071	235.7881	223.0355	243.8426	223.4541	234.2603	221.7079	242.1996
0.15	0.15	230.4004	241.0149	228.6758	248.8245	228.9519	239.3935	227.2525	247.0903
0.15	0.20	236.0913	246.3046	234.4191	253.8493	234.5427	244.5859	232.8954	252.0208
0.20	0.05	199.1163	208.4569	197.6162	215.1987	198.1634	207.3767	196.6817	214.0359
0.20	0.10	204.9445	214.0888	203.4670	220.7086	203.9063	212.9220	202.4475	219.4593
0.20	0.15	210.9229	219.8402	209.4726	226.3199	209.7939	218.5820	208.3624	224.9796
0.20	0.20	217.0573	225.7149	215.6391	232.0353	215.8316	224.3602	214.4322	230.5994

Table 6. PRE of T w.r.t. τ for real data.

Stratum	PRE (In the absence of measurement error)	PRE (In the presence of measurement error)
Case I	286.7866	284.2111
Case II	307.7331	304.693
Case III	283.4925	280.983
Case IV	323.7506	320.3699

Table 7. In the absence of non-response, PRE is observed from real data when $p_1 = p_2 = 0$.

The recommended estimator successfully mitigates the adverse effects of random non-response and measurement errors in two-phase stratified sampling. When additional information on two positively related variables is available, the advantages are evident. We anticipate the evolution of more estimators within the proposed class, allowing survey statisticians to provide even more precise estimates.

Conclusions

Our research has illuminated several critical contributions and practical applications:

The calibration technique significantly enhances the accuracy of stratum weights, leading to more precise estimates, even in the presence of minor deviations from the actual weights. The proposed estimator consistently outperforms its counterparts within specific parameter ranges, showcasing its robustness in handling measurement errors. The superior Percentage Relative Efficiency (PRE) of our proposed estimator, even in scenarios

without non-response, highlights its effectiveness in improving estimation accuracy. We've observed that the correlation coefficient and the values of p_1 and p_2 play significant roles in the performance of the estimator. The versatility of our estimation approach extends its applicability across diverse fields, including the estimation of variance in simulated data. The results obtained from simulated data are further validated through the analysis of real-world data, such as gas turbine exhaust pressure, confirming the applicability and reliability of our proposed methodology in practical scenarios.

Our study provides valuable methodologies to enhance population variance estimation, particularly in practical scenarios rife with non-response and measurement errors. The consistent and outstanding performance of our proposed estimators corroborates their effectiveness and reliability within the domain of survey statistics. Moreover, incorporating neutrosophic statistics aligns with the need to address uncertainty and imprecision in survey data, further reinforcing the effectiveness of our proposed methodology. The validation of our simulated data against real-world datasets substantiates the applicability and trustworthiness of our proposed methodology in practical, real-life scenarios.

Data availibility

Secondary data used in the manuscript is freely available from the UCI Machine Learning Repository dataset named 'Gas Turbine CO and NOx Emission Data Set.' For the above analysis, we have chosen the file '$gt_{2011}.csv$'. The data can be accessed from https://archive.ics.uci.edu/dataset/551/gas+turbine+co+and+nox+emission+data+set.

Received: 12 July 2023; Accepted: 10 November 2023
Published online: 05 February 2024

References

1. Das, A. K. Use of auxiliary information in estimating the finite population variance. *Sankhya, C* **40**, 139–148 (1978).
2. Srivastava, S. K. A class of estimators using auxiliary information for estimating finite population variance. *Sankhya C* **42**, 87–96 (1980).
3. Singh, H. P. & Solanki, R. S. A new procedure for variance estimation in simple random sampling using auxiliary information. *Statist. Papers* **54**, 479–497 (2013).
4. Ahmad, S. *et al.* Improved estimation of finite population variance using dual supplementary information under stratified random sampling. *Math. Probl. Eng.* **2022** (2022).
5. Ahmad, S. *et al.* A simulation study: Using dual ancillary variable to estimate population mean under stratified random sampling. *Plos One* **17**, e0275875 (2022).
6. Ullah, K., Hussain, Z. & Cheema, S. A. Using auxiliary information more efficiently in population variance estimation-a new family of estimators. *Statist. Comput. Interdisciplin. Res.* **2**, 1–12 (2020).
7. Aslam, I., Noor-ul Amin, M., Yasmeen, U. & Hanif, M. Memory type ratio and product estimators in stratified sampling. *J. Reliab. Statist. Stud.* 1–20 (2020).
8. Aslam, I., Noor-Ul Amin, M., Hanif, M. & Sharma, P. Memory type ratio and product estimators under ranked-based sampling schemes. *Commun. Statist. Theory Methods* **52**, 1155–1177 (2023).
9. Singh, H. P. & Karpe, N. A class of estimators using auxiliary information for estimating finite population variance in presence of measurement errors. *Commun. Statist. Theory Methods* **38**, 734–741 (2009).
10. Shukla, D., Pathak, S. & Thakur, N. S. Class (es) of factor-type estimator (s) in presence of measurement error. *J. Modern Appl. Statist. Methods* **11**, 6 (2012).
11. Misra, S., Yadav, D. K., Dipika, A. & Shukla, K. On estimation of population coefficient of variation in presence of measurement errors. *Int. J. Math. Trends Technol.* **51**, 307–311 (2017).
12. Zahid, E. & Shabbir, J. Estimation of population mean in the presence of measurement error and non response under stratified random sampling. *PloS One* **13**, e0191572 (2018).
13. Hussain, S., Ahmad, S., Akhtar, S., Javed, A. & Yasmeen, U. Estimation of finite population distribution function with dual use of auxiliary information under non-response. *Plos One* **15**, e0243584 (2020).
14. Zahid, E., Shabbir, J., Gupta, S., Onyango, R. & Saeed, S. A generalized class of estimators for sensitive variable in the presence of measurement error and non-response. *Plos One* **17**, e0261561 (2022).
15. Ahmad, S. *et al.* Estimation of finite population mean using dual auxiliary variable for non-response using simple random sampling. *Aims Math.* **793**, 4592–4613 (2022).
16. Bhushan, S., Kumar, A. & Shukla, S. On classes of robust estimators in presence of correlated measurement errors. *Measurement* **220**, 113383 (2023).
17. Bhushan, S. *et al.* New logarithmic type imputation techniques in presence of measurement errors. *Alexandria Eng. J.* **71**, 707–730 (2023).
18. Deming, W. E. & Stephan, F. F. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11**, 427–444 (1940).
19. Deville, J. C. & Särndal, C. E. Calibration estimators in survey sampling. *J. Am. Statist. Associat.* **87**, 376–382 (1992).
20. Särndal, C. E. The calibration approach in survey theory and practice. *Surv. Methodol.* **33**, 99–119 (2007).
21. Singh, G. N., Bhattacharyya, D. & Bandyopadhyay, A. A general class of calibration estimators under stratified random sampling in presence of various kinds of non-sampling errors. *Commun. Statist. Simulat. Comput.* **52**, 320–333 (2023).
22. El-Sheikh, A. A. & El-Kossaly, H. A. Calibration estimation for ratio estimators in stratified sampling for proportion allocation. *J. Progress. Res. Math.* **16**, 3199–3205 (2020).
23. Hussain, S., Ahmad, S., Saleem, M. & Akhtar, S. Finite population distribution function estimation with dual use of auxiliary information under simple and stratified random sampling. *Plos One* **15**, e0239098 (2020).
24. Hussain, S., Akhtar, S. & El-Morshedy, M. Modified estimators of finite population distribution function based on dual use of auxiliary information under stratified random sampling. *Sci. Progr.* **105**, 00368504221128486 (2022).
25. Smarandache, F. Neutrosophy: Neutrosophic Probability, Set, and Logic: Analytic Synthesis & Synthetic Analysis. 1998.
26. Aslam, M. A new sampling plan using neutrosophic process loss consideration. *Symmetry* **10**, 132 (2018).
27. Aslam, M. Neutrosophic analysis of variance: Application to university students. *Complex Intell. Syst.* **5**, 403–407 (2019).
28. Singh, S. & Joarder, A. H. Estimation of finite population variance using random non-response in survey sampling. *Metrika* **47**, 241–249 (1998).

Acknowledgements
Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R368), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions
The concept for developing the estimator and conducting empirical studies was proposed by M.K.P. Theoretical analysis, comparisons, and literature review were carried out by G.N.S., while T.Z. formulated the complete research article. Funding acquisition was managed by A.A.M., and M.S.M. contributed to the software aspect. The final manuscript was reviewed and approved by all authors.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-47234-1.

Correspondence and requests for materials should be addressed to M.K.P.