Sains Malaysiana 53(7)(2024): 1693-1702
http://doi.org/10.17576/jsm-2024-5307-16

Enhancing Precision in Population Variance Vector Estimation: A Two-Phase Sampling Approach with Multi-Auxiliary Information

(Meningkatkan Ketepatan dalam Anggaran Vektor Varians Populasi: Pendekatan Persampelan Dua Fasa dengan Maklumat Berbilang Bantu)

AMBER ASGHAR[1], AAMIR SANAULLAH[2,*], MUHAMMAD HANIF[3] & LAILA A. AL-ESSA[4]

[1]Department of Statistics, Faculty of Science and Technology, Virtual University of Pakistan, Lahore, Pakistan
[2]Department of Statistics, COMSATS University Islamabad, Lahore Campus, Pakistan
[3]Department of Statistics, National College of Business Administration & Economics, Lahore, Pakistan
[4]Department of Mathematical Sciences, College of Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Received: 11 January 2024/Accepted: 11 June 2024

ABSTRACT

To enhance precision in estimating unknown population parameters, an auxiliary variable is often used. However, in scenarios where required information on an auxiliary variable is partially or fully unavailable, two-phase sampling is commonly employed. The challenge of estimating the variance vector using multi-auxiliary variables is a less explored area in current literature. This paper addresses the estimation of vector of unknown population variances for multiple study variables by using an estimated vector of variances derived from multi-auxiliary information. This approach is particularly relevant when population variances for the multi-auxiliary variables are not known prior to the survey. The paper introduces a generalized variance and a vector of biases for the proposed multivariate estimator. Special cases of the proposed multivariate variance estimator are provided, accompanied by expressions for mean square errors. Theoretical mathematical conditions are discussed to guide the preference for the proposed estimator. Through the analysis of real-world application-based data, the applicability and efficiency of the proposed multivariate variance estimator are demonstrated, outperforming modified versions of multivariate variance estimators. Additionally, a simulation study validates the superior performance of the proposed estimator compared to its modified estimators.

Keywords: Generalized variance; multivariate estimator; regression-cum-exponential estimator; two-phase sampling; variance vector estimator

ABSTRAK

Untuk meningkatkan ketepatan dalam menganggar parameter populasi yang tidak diketahui, pemboleh ubah bantuan sering digunakan. Walau bagaimanapun, dalam senario yang mana maklumat yang diperlukan tentang pemboleh ubah bantuan sebahagian atau sepenuhnya tidak tersedia, pensampelan dua fasa biasanya digunakan. Cabaran untuk menganggar vektor varians menggunakan pemboleh ubah berbilang bantu adalah bidang yang kurang diterokai dalam kepustakaan semasa. Kertas ini menangani anggaran vektor varians populasi yang tidak diketahui untuk pelbagai pemboleh ubah kajian dengan menggunakan anggaran vektor varians yang diperoleh daripada maklumat berbilang bantu. Pendekatan ini amat relevan apabila varians populasi untuk pemboleh ubah berbilang bantu tidak diketahui sebelum tinjauan. Makalah ini memperkenalkan varians umum dan vektor bias untuk penganggar multivariat yang dicadangkan. Kes khas penganggar varians multivariat yang dicadangkan disediakan, disertakan dengan pengekspresan untuk ralat kuasa dua min. Keadaan matematik teori dibincangkan untuk membimbing keutamaan bagi penganggar yang dicadangkan. Melalui analisis data berasaskan aplikasi dunia sebenar, kebolehgunaan dan kecekapan penganggar varians multivariat yang dicadangkan ditunjukkan, mengatasi versi pengubahsuaian penganggar varians multivariat. Selain itu, kajian simulasi mengesahkan prestasi unggul penganggar yang dicadangkan berbanding penganggarnya yang diubah suai.

Kata kunci: Penganggar multivariat; penganggar regresi merangkap eksponen; penganggar vektor varians; pensampelan dua fasa; varians umum

Introduction

In recent years, survey sampling has become helpful in various sectors, including academia, healthcare, and both public and private industries. Surveys, employing both probability and non-probability sampling, are important in fields such as agriculture, industry, and healthcare. The crucial role of survey sampling in collecting data across diverse fields is indisputable.

As the general use of survey sampling grows, the demand for more advanced methods to interpret results becomes dominant. Multivariate estimation is also one of them. For example, in environmental monitoring, there are instances where auxiliary information (such as meteorological conditions) is not available for all monitoring stations or time periods. The use of a multivariate estimator (MV) without auxiliary information allows researchers to still derive estimates of air quality indicators, providing valuable insights into city-wide pollution levels. Among these methods, variance estimation emerges as the best choice for addressing the elaboration of complex survey designs. Neyman (1938) introduced a cost-effective two-phase sampling technique which is particularly valuable when collecting data on the variable of interest proves financially burdensome. For a deeper understanding of the application of two-phase sampling, refer to Breidt and Fuller (1993), Cochran (1977), Hussain et al. (2018), and Rao (1973).

Cebrián and García (1997) proposed an almost unbiased multivariate ratio-type estimator for population variance. Das and Tripathi (1978) addressed variance estimation by incorporating both population variance and the coefficient of variation (CV) of an auxiliary variable. Isaki (1983) developed ratio and regression estimators using the variance of the auxiliary variable for variance estimation.

Singh, Chandra and Singh (2003) proposed a variance estimator using multi-auxiliary variables (MAVs) and explored Srivastava and Jhajj (1980) estimators. Ahmad, Hussain and Hanif (2016) suggested a multivariate approach under successive sampling. Asghar, Sanaullah and Hanif (2018) introduced a multivariate variance (MV) estimator for variance-vector estimation using MAVs in two-phase sampling. Further, Asghar et al. (2023) proposed the multivariate ratio estimator for estimating the variance vector. Zamanzade and Al-Omari (2016) provided estimates of mean and variance by incorporating modifications to traditional ranked set sampling. Muneer et al. (2018) introduced a new ratio-cum-product exponential-type estimator for the unknown variance of a finite population. Lone, Subzar and Sharma (2021) enhanced the performance of a population variance estimator by using the supporting information.

Shahzad et al. (2021a) estimated variance using attributes of auxiliary variables, while Zaman and Bulut (2019) proposed a generalized variance approach instead of traditional ratio estimators. Additionally, Shahzad et al. (2021b) proposed estimators for population variance based on L-moments, such as L-mean, L-standard deviation, and L-coefficient of variation.

Various researchers have proposed different estimators for the estimation of population variance under the assumption that the population variance of auxiliary variables is known prior to a survey. Notable works include those by Ahmad, Hussain and Hanif (2016), Arcos and Rueda (1997), Asghar et al. (2023), Niaz et al. (2022), Sanaullah et al. (2020), and Zaman and Bulut (2019). However, real-life situations often arise where population variances are not available before a survey. In such cases, two-phase sampling becomes essential to estimate the unknown population variance of auxiliary variables.

Researchers, such as Abu-Dayyeh and Ahmed (2005), have addressed this issue in the context of estimating the variance of a single study variable. Additionally, Sanaullah, Hanif and Asghar (2016) introduced generalized exponential-type ratio and product estimators for estimating variance when the mean auxiliary variable is unknown prior to the survey under two-phase sampling. More recently, Abid et al. (2020) proposed a ratio estimator for robust measures of population variance and compared it with competing estimators using quantiles of auxiliary variables, concluding its robust performance even in the presence of outliers.

Motivated by the studies discussed in the earlier text, this paper introduces a novel multivariate regression-cum-exponential (MRCE) type estimator for scenarios where information about the parameter(s) of auxiliary variables is not available. While usual regression and multiple regression estimators perform well in symmetric and linear situations, practical scenarios often exhibit non-symmetric and skewed distributions. In such cases, regression-cum-exponential type estimators are expected to yield superior results compared to simple regression estimators.

Materials, Notations and Sampling Methodology

Consider a finite population comprising N units. Let Y_j represent the j-th study variable, and X_k denote the k-th auxiliary variable, where j ranges from 1 to m and k ranges from 1 to n. Utilizing a two-phase sampling design, a first-phase sample size n_1 is initially selected, followed by a second-phase sample size n_2 chosen as a function of n_1 (i.e., $n_2 = f(n_1)$). To comprehend the properties of an estimator, certain essential expectations are demonstrated by examining the following expressions:

$$E\left(\bar{e}_{x_{(1)k}}\right) = E\left(\bar{e}_{x_{(2)k}}\right) = E\left(\varepsilon_{x_{(2)k}}\right) = E\left(\varepsilon_{y_2}\right) = E\left(\varepsilon_{y_{(1)j}}\right) = 0$$

$$E_2\left(\varepsilon^2_{y_{(2)}}\right) = \gamma_2 A_y, E_2\left(\varepsilon^2_{x_{(1)}}\right) = \gamma_2 A_x,$$

$$E_1E_{2/1}\left(\varepsilon_{y_2}\varepsilon_x\right) = \gamma_1 A_{yx}, E_1E_{2/1}\left(\bar{e}_{x_{(1)k}}\varepsilon_{y_2}\right) = \gamma_1 A_{yx_d},$$

$$E_1\left(\bar{e}^2_{x_{(1)k}}\right) = \gamma_1 A_{x_{d(1)k}}, \quad E_1(\bar{e}_{x_{(1)k}}\bar{e}_{x_{(1)l}}) = \gamma_1 A_{x_k x_l}, \text{ where } k \neq l,$$

where E_1 and E_2 denote the expectations over the first and second phase, and $\gamma_1 = \frac{1}{n_1}$, & $\gamma_2 = \frac{1}{n_2}$.

For estimation of vector of population variances, let us consider the notations and expectations given by, and

$$\bar{D}_x = \left[\mathrm{e}_{x_{(1)1}} \quad \mathrm{e}_{x_{(1)2}} \quad . \quad . \quad \mathrm{e}_{x(1)l}\right],$$

$$\Delta_x = \left[\varepsilon_{x_{(1)1}} \quad \varepsilon_{x_{(1)2}} \quad . \quad . \quad \varepsilon_{x_{(1)l}}\right],$$

$$\Delta_y = \left[\varepsilon_{y_{(2)1}} \quad \varepsilon_{y_{(2)2}} \quad . \quad . \quad \varepsilon_{y_{(2)\mathrm{m}}}\right],$$

$$\bar{D}_{x_1} = \left[\left(\mathrm{e}_{x_{(1)1}} - \mathrm{e}_{x_{(2)1}}\right) \quad \left(\mathrm{e}_{x_{(1)2}} - \mathrm{e}_{x_{(2)2}}\right) \quad . \quad . \quad \left(\mathrm{e}_{x_{(1)l}} - \mathrm{e}_{x_{(2)l}}\right)\right],$$

$$\Delta_{x_1} = \left[\left(\varepsilon_{x_{(1)1}} - \varepsilon_{x_{(2)1}}\right) \quad \left(\varepsilon_{x_{(1)2}} - \varepsilon_{x_{(2)2}}\right) \quad . \quad . \quad \left(\varepsilon_{x_{(1)l}} - \varepsilon_{x_{(2)l}}\right)\right],$$

$$E_1E_{2/1}\left(\bar{D}_x{}'\bar{D}_x\right) = \gamma_1 \Sigma_{x_{d(l\times l)}}, E_1E_{2/1}\left(\Delta_x{}'\Delta_y\right) = \gamma_1 \Sigma_{xy_{(l\times m)}},$$

$$E_1E_{2/1}\left(\Delta_x{}'\Delta_x\right) = \gamma_1 \Sigma_{x_{(l\times l)}}, E_1E_{2/1}\left(\Delta_y{}'\Delta_y\right) = \gamma_2 \Sigma_{y_{(\mathrm{m}\times m)}},$$

and $E_1E_{2/1}\left(\Delta_y{}'\bar{D}_x\right) = \gamma_1 \Sigma_{yx_{d(\mathrm{m}\times l)}}, E_1E_{2/1}\left(\bar{D}_x{}'\Delta_y\right)$

$$= \gamma_1 \Sigma_{x_d y_{(l\times m)}}.$$

Let S^2_y be the population variance, and its usual unbiased sample variance estimator vector, is given by

$$t_0 = \left[t_{0j}\right]_{(1\times m)}, \text{ where, } \quad t_{0j} = s^2_{y_{j(2)}}, \qquad j = 1, 2, \ldots, m. \quad (1)$$

The variance of the modified unbiased sample variance vector is defined as,

$$\Sigma_{t_0} = \gamma_2 \Sigma_{y_{(\mathrm{m}\times m)}}.$$

The Isaki (1983) estimator is modified into a multivariate regression (MR) estimator, and form of the modified estimator, is given by,

$$t_{reg} = \left[t_{regj}\right]_{(1\times m)}, \text{ where, } t_{regj}$$

$$= s^2_{y_{\mathrm{j}(2)}} + \sum_{k=1}^{l} \delta_{kj}\left(s^2_{x_{(1)k}} - s^2_{x_{(2)\mathrm{k}}}\right). \quad (2)$$

The expression of Generalized variance of t_{reg}, is given by,

$$\Sigma_{t_{reg}} = S'S\left[\gamma_2 \Sigma_{y_{(\mathrm{m}\times m)}} - (\gamma_2 - \gamma_1)\Sigma_{yx_{(\mathrm{m}\times l)}} \Sigma^{-1}_{x_{(l\times l)}} \Sigma_{xy_{(l\times m)}}\right].$$

Following Shabbir and Gupta (2015), a MR type estimator using sample means of auxiliary information is modified, and it is given by,

$$t_{rg} = \left[t_{rgj}\right]_{(1\times m)}, \text{ where, } t_{rgj}$$

$$= s^2_{y_{\mathrm{j}(2)}} + \sum_{k=1}^{l} \pi_{kj}\left(\bar{x}_{(1)\mathrm{k}} - \bar{x}_{(2)\mathrm{k}}\right). \quad (3)$$

The expression of Generalized variance of t_{rg} is shown by,

$$\Sigma_{t_{rg}} = S'S\left[\gamma_2 \Sigma_{y_{(\mathrm{m}\times m)}} - (\gamma_2 - \gamma_1)\Sigma_{yx_{d(\mathrm{m}\times l)}} \Sigma^{-1}_{x_{d(l\times l)}} \Sigma_{x_d y_{(l\times m)}}\right].$$

Following Sanaullah, Hanif and Asghar (2016), a MRCE type estimator is modified for estimating a vector of population variances, and it is given with a form given by, where,

$$t_a = \left[t_a\right]_{(1\times m)}, \text{ where, } t_{aj} = \left(s^2_{y_{\mathrm{j}(2)}} + \sum_{k=1}^{l} \pi_{kj}\left(\bar{x}_{(1)\mathrm{k}} - \bar{x}_{(2)\mathrm{k}}\right)\right)$$

$$\left(\exp \sum_{k=1}^{l} d_{kj}\left(\frac{\bar{x}_{(1)\mathrm{k}} - \bar{x}_{(2)\mathrm{k}}}{\bar{x}_{(1)\mathrm{k}} + \bar{x}_{(2)\mathrm{k}}}\right)\right). \quad (4)$$

The Generalized variance of t_a is given by,

$$\Sigma_{t_a} = S'S\left[\gamma_2 \Sigma_{y_{(\mathrm{m}\times m)}} - (\gamma_2 - \gamma_1)\Sigma_{yx_{d(\mathrm{m}\times l)}} \Sigma^{-1}_{x_{d(l\times l)}} \Sigma_{x_d y_{(l\times m)}}\right].$$

PROPOSED METHODOLOGY FOR ESTIMATING THE VECTOR OF VARIANCES

In this section, a generalized MRCE estimator under two-phase sampling is proposed for estimating a vector of population variances. Form of the proposed MV estimator is given by,

$$t_p = \left[t_{pj}\right]_{(1xm)}, \quad j = 1,2,3,...,m,$$

where, $t_{pj} = \left(s^2_{y_{j(2)}} + \sum_{k=1}^{l} \delta_{kj}\left(s^2_{x(1)\mathrm{k}} - s^2_{x(2)\mathrm{k}}\right)\right) \exp$

$$\left(\sum_{k=1}^{l} b_{kj}\left(\frac{s^2_{x(1)\mathrm{k}} - s^2_{x(2)\mathrm{k}}}{s^2_{x(1)\mathrm{k}} + s^2_{x(2)\mathrm{k}}}\right)\right). \tag{5}$$

It is clear that positive values of b_{kj} produce different families of multivariate exponential (ME) ratio estimators, and negative values of b_{kj} will give different families of ME product estimators. δ_{kj} is assumed to be unknown and needs to be estimated and emphasized its optimal value.

DERIVATION OF THE VECTOR OF THE BIAS, AND THE GENERALIZED VARIANCE

To determine the generalized variance and bias vector, the proposed estimator, as defined in Equation (5), incorporates sampling errors. At the start, focus on the *j*-th estimator of this proposed estimator, expressed as follows,

$$t_{pj} = \left(\mathrm{S}^2_{y_{j(2)}}\left(1+\varepsilon_{y_{j(2)}}\right) + \sum_{k=1}^{l}\delta_{kj}\left\{\mathrm{S}^2_{x_k}\left(1+\varepsilon_{x_{(1)k}}\right) - \mathrm{S}^2_{x_k}\left(1+\varepsilon_{x_{(2)k}}\right)\right\}\right)$$
$$\left(\exp\left(\sum_{k=1}^{l} b_{kj}\left[\frac{\mathrm{S}^2_{x_k}\left(1+\varepsilon_{x_{(1)\mathrm{k}}}\right) - \mathrm{S}^2_{x_k}\left(1+\varepsilon_{x_{(2)\mathrm{k}}}\right)}{\mathrm{S}^2_{x_k}\left(1+\varepsilon_{x_{(1)\mathrm{k}}}\right) + \mathrm{S}^2_{x_k}\left(1+\varepsilon_{x_{(2)_k}}\right)}\right]\right)\right),$$

or

$$t_{pj} = \left(\mathrm{S}^2_{y_{j(2)}}\left(1+\varepsilon_{y_{j(2)}}\right) + \sum_{k=1}^{l}\delta_{kj}\left\{\varepsilon_{x_{(1)k}} - \varepsilon_{x_{(2)k}}\right\}\right)$$
$$\exp\left(\sum_{k=1}^{l} b_{kj}\left[\frac{\left(\varepsilon_{x_{(1)\mathrm{k}}} - \varepsilon_{x_{(2)\mathrm{k}}}\right)}{2}\left(1+\frac{\left(\varepsilon_{x_{(1)\mathrm{k}}} + \varepsilon_{x_{(2)\mathrm{k}}}\right)}{2}\right)^{-1}\right]\right). \tag{6}$$

Retaining the terms up to the order $O(n^{-1})$, the proposed estimator is given as,

$$t_{pj} = \left(\mathrm{S}^2_{y_{j(2)}}\left(1+\varepsilon_{y_{j(2)}}\right) + \sum_{k=1}^{l}\delta_{kj}\left\{\varepsilon_{x_{(1)k}} - \varepsilon_{x_{(2)k}}\right\}\right) \exp$$
$$\left(\sum_{k=1}^{l} b_{kj}\left[\frac{\left(\varepsilon_{x_{(1)\mathrm{k}}} - \varepsilon_{x_{(2)\mathrm{k}}}\right)}{2}\left(1 - \frac{\left(\varepsilon_{x_{(1)\mathrm{k}}} + \varepsilon_{x_{(2)\mathrm{k}}}\right)}{2} + \frac{\left(\varepsilon_{x_{(1)\mathrm{k}}} + \varepsilon_{x_{(2)\mathrm{k}}}\right)^2}{4} \pm ...\right)\right]\right), \tag{7}$$

After some simplification, the equation is obtained as,

$$t_{Pj} = \mathrm{S}^2_{y_{j(2)}}$$
$$\begin{pmatrix} 1 + \varepsilon_{y_{j(2)}} + \frac{1}{2}\sum_{k=1}^{l}\left(2\delta_{kj}\frac{S^2_{x_k}}{\mathrm{S}^2_{y_{j(2)}}} + b_{kj}\right)\left(\varepsilon_{x_{(1)k}} - \varepsilon_{x_{(2)k}}\right) \\ -\frac{1}{8}\sum_{k=1}^{l} b^2_{kj}\left(\varepsilon_{x_{(1)k}} - \varepsilon_{x_{(2)k}}\right)^2 + \frac{1}{2}\sum_{k=1}^{l}\left(b_{kj}\right)\varepsilon_{y_{j(2)}}\left(\varepsilon_{x_{(1)k}} - \varepsilon_{x_{(2)k}}\right) \\ -\frac{1}{4}\sum_{k=1}^{l} b_{kj}\left(\varepsilon^2_{x_{(1)k}} - \varepsilon^2_{x_{(2)k}}\right) + \frac{1}{2}\sum_{k=1}^{l}\frac{S^2_{x_k}}{\mathrm{S}^2_{y_{j(2)}}}\delta_{kj} b_{kj}\left(\varepsilon_{x_{(1)k}} - \varepsilon_{x_{(2)k}}\right)\left(\varepsilon_{x_{(1)k}} - \varepsilon_{x_{(2)k}}\right) \end{pmatrix} \tag{8}$$

Applying expectations, a vector of the bias expression for *j*-th estimator is given by,

$$Bias\left(t_p\right)_{(1\times m)} = \mathrm{S}^2_{y_{j(2)}}$$
$$\begin{pmatrix} \frac{1}{8}(\gamma_2 - \gamma_1) A_{x_{(l\times l)}}\left[b^2_{kj}\right]_{(l\times m)} + \frac{1}{4}(\gamma_2 - \gamma_1) A_{x_{(l\times l)}}\left[b_{kj}\right]_{(l\times m)} \\ -\frac{1}{2}(\gamma_2 - \gamma_1) A_{yx_{(l\times l)}}\left[b_{kj}\right]_{(l\times m)} + \frac{1}{2}(\gamma_2 - \gamma_1) A_{x_{(l\times l)}}\left[\delta_{kj}\frac{S^2_{x_k}}{\mathrm{S}^2_{y_{j(2)}}} b_{kj}\right]_{(l\times m)} \end{pmatrix}. \tag{9}$$

Consider Equation (8) to proceed as,

$$\boldsymbol{\Sigma}_{t_{p(\mathrm{m}\times m)}} = E_1 E_{2/1}\left(\mathbf{t}_{p(l\times m)} - \mathrm{S}_{(l\times m)}\right)'\left(\mathbf{t}_{p(l\times m)} - \mathrm{S}_{(l\times m)}\right) = S^4_y E_1 E_{2/1}$$
$$\left[\left(\Delta_{y_{(1\times m)}} + \frac{1}{2}\Delta_{x_{(l\times l)}}\Psi_{(l\times m)}\right)'\left(\Delta_{y_{(1\times m)}} + \frac{1}{2}\Delta_{x_{(l\times l)}}\Psi_{(l\times m)}\right)\right], \tag{10}$$

where, $\Psi_{(l\times m)} = \left(2\delta_{kj}\phi + b_{kj}\right)_{(l\times m)}$ $(k = 1,2,...,l,\ j = 1,2,...,m)$,

and $\phi = \dfrac{S^2_{x_k}}{\mathrm{S}^2_{y_{j(2)}}}$.

Alternatively, Equation (10) can take another form, given by:

$$\Sigma_{t_{p(\mathrm{m}\times m)}} = S'S$$
$$\begin{pmatrix} \gamma_2 \Sigma_{y_{(\mathrm{m}\times m)}} + \frac{1}{2}(\gamma_1 - \gamma_2)\Sigma_{yx_{(m\times l)}}\Psi_{(l\times m)} \\ +\frac{1}{2}(\gamma_1 - \gamma_2)\Psi'_{(l\times m)}\Sigma_{yx_{(m\times l)}} + \frac{1}{4}(\gamma_2 - \gamma_1)\Psi'_{(\mathrm{m}\times l)}\Sigma_{x_{(l\times l)}}\Psi_{(l\times m)} \end{pmatrix}. \tag{11}$$

The optimum value of Ψ are obtained by differentiating Equation (11) w.r.t Ψ, and equating the first derivative to zero. The optimum value is given by,

$$\Psi_{opt(l\times m)} = 2\sum\nolimits^{-1}_{x_{(l\times l)}} \sum_{yx_{(l\times m)}}.$$

Put the optimum value of Ψ in Equation (11), the minimum Generalized variance of t_p is obtained, and it is given by

$$\min \sum_{t_{p(\mathrm{m}\times m)}} = S'S\left(\gamma_2 \sum_{y_{(\mathrm{m}\times m)}} - \gamma_1 \sum_{yx_{(\mathrm{m}\times l)}} \sum\nolimits^{-1}_{x_{(l\times l)}} \sum_{yx_{(l\times m)}}\right). \quad (12)$$

Remark 1

It is observed that considering $b_{kj} = 0$ in Equation (5) the MR estimator may be obtained for the no information case using MAVs. The vector of the bias and the generalized variance may be obtained directly from Equations (9) and (11).

$$t_r = \left[t_{rj}\right]_{(1\times m)}, \quad j = 1,2,3,...,m,$$

where, $t_{rj} = \left(s^2_{y_{j(2)}} + \sum_{k=1}^{l} \delta_{kj}(s^2_{x(1)\mathrm{k}} - s^2_{x(2)\mathrm{k}})\right)$.

Remark 2

Similarly, it is also noted that the ME estimator may be obtained by putting $\delta_{kj} = 0$ directly in Equation (5) and its bias and generalized variance may be obtained from Equations (9) and (11) for the situation when the required information about the parameters or the function of parameters, such as the population means, variances, correlation, coefficient of variation etcetera of the MAVs is not available.

$$t_e = \left[t_{ej}\right]_{(1\times m)}, \quad j = 1,2,3,...,m,$$

where, $t_{ej} = \exp\left(\sum_{k=1}^{l} b_{kj}\left(\frac{s^2_{x(1)\mathrm{k}} - s^2_{x(2)\mathrm{k}}}{s^2_{x(1)\mathrm{k}} + s^2_{x(2)\mathrm{k}}}\right)\right)$.

Remark 3

Observing that for $j = 1$, and $k = 1, 2, 3, \ldots, l$, is considered in Equation (5), a generalized univariate regression-cum-exponential estimator can be derived. This setting refers to situations where no information is available, and the resulting estimator can be expressed as,

$$t_{pu} = \left[t_{puj}\right]_{(1\times 1)}, \; j = 1,2,3,...,m,$$

where, $t_{pu_1} = \left(s^2_{y_{(2)}} + \sum_{k=1}^{l} \delta_k (s^2_{x(1)\mathrm{k}} - s^2_{x(2)\mathrm{k}})\right) \exp$

$$\left(\sum_{k=1}^{l} b_k \left(\frac{s^2_{x(1)\mathrm{k}} - s^2_{x(2)\mathrm{k}}}{s^2_{x(1)\mathrm{k}} + s^2_{x(2)\mathrm{k}}}\right)\right).$$

Under the same conditions, one may obtain the bias and generalized variance for the generalized univariate estimator from Equations (9) and (11).

RELATIVE PERFORMANCE OF THE PROPOSED ESTIMATOR

To assess the effectiveness of the suggested MREV estimator, a theoretical comparison is conducted, evaluating its performance in a univariate context against various alternative modified estimators.

The proposed estimator t_p will perform more efficiently than usual sample variance estimator t_0 iff,

$$\frac{A_{yx}}{A_x} > \frac{\delta\phi}{2} + \frac{1}{4}. \quad (13)$$

The proposed estimator t_p performs better than Isaki's (1983) regression estimator t_{reg} when

$$\frac{A_{yx}}{A_x} > \delta\phi + \frac{1}{4}. \quad (14)$$

The proposed estimator t_p performs better than Shabbir and Gupta (2015) regression estimator t_{rg} when following condition is met,

$$\left(\sqrt{A_x}\Psi - 2\frac{A_{yx}}{\sqrt{A_x}}\right)^2 < c,$$

$$\frac{A_{yx}}{A_x} > \phi\delta + \frac{1}{2}\left(1 - \frac{\sqrt{c}}{\sqrt{A_x}}\right), \text{ or } \frac{A_{yx}}{A_x} > -\phi\delta + \frac{1}{2}\left(1 + \sqrt{\frac{c}{A_x}}\right),$$

where, $c = \left(\pi^2\phi_1^2 A_{x_d} - 8A_{yx_d}\pi\phi_1 + \left(\frac{2A_{yx}}{\sqrt{A_x}}\right)^2\right)$. (15)

The proposed estimator t_p performs better than Sanaullah, Hanif and Asghar (2016) regression-cum-exponential estimator t_a, when,

$$\frac{A_{yx}}{A_x} > -\phi\delta - \frac{1}{2}\left[\frac{\sqrt{d}}{\sqrt{A_x}} - 1\right], \text{ or } \frac{A_{yx}}{A_x} > \phi\delta - \frac{1}{2}\left[\frac{\sqrt{d}}{\sqrt{A_x}} - 1\right],$$

$$\text{where, } d = \beta A_{x_d}\left(\beta - 4\frac{A_{yx_d}}{A_{x_d}}\right) + \left(\frac{2A_{yx}}{\sqrt{A_x}}\right)^2. \quad (16)$$

Results and Discussion

In this study, the MRCE is utilized, demonstrating its application with real-life data extracted from Canadian climate records, as published by the National Oceanic and Atmospheric Administration (NOAA). The dataset focuses on daily weather records for the month of May 2017, collected from 37,247 different weather stations across Canada.

To contour the analysis, transform the data into weekly summaries, specifically for the weeks of May 08-14, 2017; May 15-21, 2017; and May 22-28, 2017. To ensure data wholeness, stations with incomplete records for a full week are excluded from the analysis. Only weeks with recorded data for the entire week are retained for further investigation.

For the auxiliary variables, monthly temperature data for the last three years (2016, 2015, and 2013) is considered. Stations lacking temperature records for the entire week in these years are excluded. As a result, 704 stations are retained with complete temperature data for the full week for our analysis. To address the missing values and facilitate variable transformations, the tidyverse Package (2016) is used. The study variable is defined as the Temperature Averages (TAVGs) for the three weeks of May 2017, denoted as i =1, 2, 3. The auxiliary variables for the previous three years (2016, 2015, and 2013) consist of monthly TAVGs, also denoted as i =1, 2, 3.

A finite population approach is used to model the variance and covariance of weekly temperatures across Canada. A comprehensive description of population characteristics is presented in Table A1 in the appendix. In this study, a two-phase sampling strategy is employed for estimating the population variance. Initially, a sample of size n_1 = 211 is selected from the population, and relevant calculations are performed based on this first-phase sample. Subsequently, using a simple random sampling without replacement, a second-phase sample of size n_2 = 105 is drawn from the previously selected first-phase sample. For both the first-phase and second-phase samples, key statistics are computed. These statistics are then integrated into the proposed multivariate (MV) estimator, as well as the existing MV estimators, to derive estimates for the population variance. This entire process is simulated 1000 times to ensure robustness in the estimation.

Finally, the determinants for the generalized variance of each MV estimator are calculated. The results are presented in Table 1, setting the corresponding values for both the proposed and modified existing estimators. To address the missing-ness in the data where it is present, mice R statistical package (2011) is used. The Broom (2016) package for replicating two-phase sampling procedure is adopted, so as to take first-phase and second phase samples in a faster and more efficient way.

Table 2 shows the determinants of the proposed, and existing estimators. From Table 2, the proposed MV estimator is attaining the smaller value of determinant of its generalized variance, where the determinants for each of the mentioned existing estimators are larger. The percent relative efficiency (PRE) value for each of the mentioned estimators including the proposed estimator is also computed. Here, PRE values also indicate that the proposed MV estimator is better in performance as compared to the performance of mentioned existing estimators. Tables 1 and 2 clearly indicate that the empirical results are in favor of the proposed MV estimator.

Simulation Results and Discussion

A simulation is performed to assess the performance of the proposed MV, along with other mentioned estimators, for estimating population variance under a two-phase sampling design. Consider the three study variables (Y_1, Y_2, Y_3) and the three auxiliary variables X_1, X_2, X_3) for multivariate estimation. The auxiliary variables are generated following the mechanism defined as follow:

$$X_1 \sim N(12, 3.3); X_2 \sim X_1^* \rho + \sqrt{1-\rho^2} * N(12, 2.5);$$

$$X_3 \sim X_2^* \rho + \sqrt{1-\rho^2} * N(15, 2.6),$$

where, $\rho = 0.9$.

The study variables are simulated using the mechanism given as:

$$Y_{ij} = \sum_{i=1}^{k} \alpha_i X_i^2 + \varepsilon, \quad \text{where, } \varepsilon \sim N(0,1).$$

The three study variables with three auxiliary variables under the discussed model are distinct by the equations:

$$Y_1 = 1.7X_1^2 + 1.3X_2^2 + 1.6X_3^2 + \varepsilon,$$

$$Y_2 = 1.2X_1^2 + 1.5X_2^2 + 1.8X_3^2 + \varepsilon,$$

$$Y_3 = 1.4X_1^2 + 1.5X_2^2 + 1.2X_3^2 + \varepsilon.$$

The statistics computed from both the first-phase and second-phase samples are then incorporated into the proposed estimator and the specified existing estimators to estimate the population variance. This simulation process is repeated 10000 times to compute the generalized variance for the proposed estimator and the mentioned existing estimators. The simulation results, including the generalized variance and the PRE values, are presented in Tables 3 and 4, respectively.

In Table 3, a value on the main diagonal of a generalized variance is a MSE of the estimator corresponding to the row and the column. Tables 3 and 4 show that the proposed estimator is more efficient than the considered estimators because a lower MSE indicates that the estimator tends to be closer to the true value of the parameter being estimated. The conclusion based on the simulation indicates that the proposed estimator remains consistent with the conclusion drawn through the real-life data.

TABLE 1. Generalized variance of the proposed and existing estimators

Generalized variance of $\mathbf{t}_p$	$\begin{pmatrix} 354.46678 & 52.91568 & 50.98563 \\ 52.91568 & 51.92468 & 37.58348 \\ 50.98563 & 37.58348 & 71.66995 \end{pmatrix}$
Generalized variance of $\mathbf{t}_0$	$\begin{pmatrix} 573.5104 & 157.29547 & 141.02297 \\ 157.2955 & 92.34649 & 74.25925 \\ 141.0230 & 74.25925 & 96.01242 \end{pmatrix}$
Generalized variance of $\mathbf{t}_{reg}$	$\begin{pmatrix} 364.71483 & 29.39202 & 24.90685 \\ 29.39202 & 62.93120 & 49.83716 \\ 24.90685 & 49.83716 & 75.40119 \end{pmatrix}$
Generalized variance of $\mathbf{t}_{rg}$	$\begin{pmatrix} 618.0714 & 172.21395 & 156.89180 \\ 172.2140 & 98.04573 & 80.48775 \\ 156.8918 & 80.48775 & 102.89652 \end{pmatrix}$
Generalized variance of $\mathbf{t}_a$	$\begin{pmatrix} 597.2441 & 168.76185 & 154.09151 \\ 168.7618 & 97.50908 & 80.23343 \\ 154.0915 & 80.23343 & 102.76829 \end{pmatrix}$

TABLE 2. Determinants of the generalized variance of the proposed and existing estimators

Estimator	$\mathbf{t}_p$	$\mathbf{t}_0$	$\mathbf{t}_{reg}$	$\mathbf{t}_{rg}$	$\mathbf{t}_a$
Determinant	1004805	685570.9	793535.9	1115737	1070918
PRE	146.5647	100	126.6238	90.0575	93.82655

TABLE 3. Simulation based on generalized variance of the proposed and existing estimators

Generalized variance of $\mathbf{t}_p$	$\begin{pmatrix} 78278828 & 79774707 & 61232339 \\ 79774707 & 81814920 & 62339927 \\ 61232339 & 62339927 & 48054423 \end{pmatrix}$
Generalized variance of $\mathbf{t}_0$	$\begin{pmatrix} 160429771 & 163784762 & 125783555 \\ 163784762 & 167725477 & 128352965 \\ 125783555 & 128352965 & 98774953 \end{pmatrix}$
Generalized variance of $\mathbf{t}_{reg}$	$\begin{pmatrix} 160420103 & 163775008 & 125774928 \\ 163775008 & 167715614 & 128344245 \\ 125774928 & 128344245 & 98767358 \end{pmatrix}$
Generalized variance of $\mathbf{t}_{rg}$	$\begin{pmatrix} 150722369 & 154107242 & 118140599 \\ 154107242 & 158089118 & 120732127 \\ 118140599 & 120732127 & 92757288 \end{pmatrix}$
Generalized variance of $\mathbf{t}_a$	$\begin{pmatrix} 160435404 & 163790189 & 125788487 \\ 163790189 & 167730726 & 128357699 \\ 125788487 & 128357699 & 98779173 \end{pmatrix}$

TABLE 4. Determinants of the proposed and existing estimators

Estimator	$\mathbf{t}_p$	$\mathbf{t}_0$	$\mathbf{t}_{reg}$	$\mathbf{t}_{rg}$	$\mathbf{t}_a$
Determinant	6.009 × 1018	1.226 × 1019	1.2258 × 1019	1.161 × 1019	1.2256 × 1019
PRE	204.0389	100	100.0186	105.5396	100.0306

CONCLUSIONS

In conclusion, this paper contributes to the field of multivariate variance estimation by introducing a novel approach that addresses the challenges of utilizing auxiliary information when the required parameters of the auxiliary variables are not readily available. In this study, the MSEs of each estimator, including the proposed estimator and competitor estimators, are computed using real-life application data (Tables 1 & 2) and simulated data (Tables 3 & 4). From these tables, it is evident that the proposed estimators have smaller MSEs than those of the competitor estimators. Therefore, it can be concluded that the proposed estimator is more efficient based on achieving smaller MSEs. The proposed MV estimator shows superior performance, as demonstrated by both real-life data analysis and simulation studies. This work lays the foundation for further advancements in precision-enhancing techniques for population parameter estimation.

ACKNOWLEDGMENTS

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R443), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

Abid, M., Sherwani, K.A.R., Tahir, M., Nazir, Z.H. & Riaz, M. 2020. An improved and robust class of variance estimator. *Scientia Arania* 28(6): 3589-3601.

Abu-Dayyeh, W. & Ahmed, M. 2005. Ratio and regression estimators for the variance under two-phase sampling. *International Journal of Statistical Sciences* 4: 49-56.

Ahmad, Z., Hussain I. & Hanif, M. 2016. Estimation of finite population variance in successive sampling using multi-auxiliary variables. *Communication in Statistics-Theory and Methods* 45(3): 553-565.

Asghar, A., Sanaullah, A. & Hanif, M. 2018. A multivariate regression-cum exponential estimator for population variance vector in two phase sampling. *Journal of King Saud University-Science* 30: 223-228.

Asghar, A., Sanaullah, A., Abbasi, A.M. & Hanif, M. 2023. Advancing sampling techniques: Multivariate ratio estimation for variance vector in two-phase sampling. *Bulletin of Business and Economics* 12(3): 473-484.

Breidt, F.J. & Fuller, W.A. 1993. Regression weighting for multipurpose samplings. *Sankhyā* 55: 297-309.

Cebrián, A.A. & García, M.R. 1997. Variance estimation using auxiliary information an almost unbiased multivariate ratio estimator. *Metrika* 45: 171-178.

Cochran, W.G. 1977. *Sampling Techniques.* New York*:* John Wiley & Sons.

Das, A.K. & Tripathi, T.P. 1978. Use of auxiliary information in estimating the finite population variance. *Sankhya* 40: 139-148.

Hussain, S., Song, L., Ahmad, S. & Riaz, M. 2018. On auxiliary information based improved EWMA median control charts. *Scientia Iranica* 25(2): 954-982.

Isaki, C. 1983. Variance estimation using auxiliary information. *Journal of the American Statistical Association* 78: 117-123.

Lone, S.A., Subzar, M. & Sharma, A. 2021. Enhanced estimators of population variance with the use of supplementary information in survey sampling. *Mathematical Problems in Engineering* 2021: 9931217.

Muneer, S., Khalila, A., Shabbirb, J. & Narjisb, G. 2018. A new improved ratio-product type exponential estimator of finite population variance using auxiliary information. *Journal of Statistical Computation and Simulation* 88(16): 3179-3192.

Neyman, J. 1938. Contribution to the theory of sampling human. *Journal of the American Statistical Association* 33(201): 101-116.

Niaz, I., Sanaullah, A., Saleem, I. & Shabbir, J. 2022. An improved efficient class of estimators for the population variance. *Concurrency and Computation: Practice and Experience* 34(4): e6620.

Rao, J. 1973. On double sampling for stratification and analytical surveys. *Biometrika* 60: 125-133.

Sanaullah, A., Hanif, M. & Asghar, A. 2016. Generalized exponential estimators for population variance under two-phase sampling. *International Journal Applied and Computational Mathematics* 2: 75-84.

Sanaullah, A., Niaz, I., Shabbir, J. & Ehsan, I. 2020. A class of hybrid type estimators for variance of a finite population in simple random sampling. C*ommunications in Statistics - Simulation and Computation* 51(10): 5609-5619.

Shabbir, J. & Gupta, S. 2015. A note on generalized exponential type estimator for population variance in survey sampling. *Revista Colombiana de Estadística* 38(2): 385-397.

Shahzad, U., Ahmad, I., Almanjahie, I.M., Al-Noor, N.MH. & Hanif, M. 2021a. A novel family of variance estimators based on L-moments and calibration approach under stratified random sampling. *Communications in Statistics - Simulation and Computation* 52(8): 3782-3795.

Shahzad, U., Ahmad, I., Almanjahie, I.M., Koyuncu, N. & Hanif, M. 2021b. Variance estimation based on L-moments and auxiliary information. *Mathematical Population Studies* 29(1): 31-46.

Singh, H.P., Chandra, P. & Singh, S. 2003. Variance estimation using multi-auxiliary information for random non-response in survey sampling. *Statistica* 63(1): 23-40.

Srivastava, S.K. & Jhajj, H.S. 1980. Class of estimator using auxiliary information for estimating finite population variance. *Sankhya* 42: 87-96.

Zaman, T. & Bulut, H. 2019. Modified regression estimators using robust regression methods and covariance matrices in stratified random sampling. *Communications in Statistics - Theory and Methods* 49(14): 3407-3420.

Zamanzade, E. & Al-Omari, A.I. 2016. New ranked set sampling for estimating the population mean and variance. *Hacettepe Journal of Mathematics and Statistic* 45(6): 1891-1905.

*Corresponding author; email: chaamirsanaullah@yahoo.com

DESCRIPTION OF THE REAL-LIFE DATA

TABLE A1. Details of variables for population

Population	Y_1 Week-1 (average temp. May 08-14, 2017)	Y_2 Week-2 (average temp. May 15-21, 2017)	Y_3 Week-3 (average temp. May 22-28, 2017)	X_1 (Year-2016 (average temp. for May)	X_2 Year-2015 (average temp. for May)	X_3 Year-2013 (average temp. for May)

TABLE A2. Variance-covariance matrix

Population	Y_1	Y_2	Y_3	X_1	X_2	X_3
Y_1	16.4344	11.7119	11.2634	5.0995	2.0234	0.5231
Y_2	11.7119	10.0879	9.8238	1.3889	2.50897	0.7491
Y_3	11.2634	9.8238	9.7610	1.3933	3.0050	0.9370
X_1	5.0996	1.3889	1.3933	7.4839	0.0406	-0.0538
X_2	2.0234	2.50897	3.0050	0.0406	4.6808	0.0080
X_3	0.5231	0.7491	0.9370	-0.0538	0.0080	2.0420

TABLE A3. Correlation matrix

Population	Y_1	Y_2	Y_3	X_1	X_2	X_3
Y_1	1.0000	0.8869	0.7654	0.9026	0.8528	0.9122
Y_2	0.8869	1.0000	0.7535	0.8397	0.7789	0.8513
Y_3	0.7654	0.7535	1.0000	0.8335	0.8118	0.8193
X_1	0.9026	0.8397	0.8335	1.0000	0.9421	-0.9453
X_2	0.8528	0.7789	0.8118	0.9421	1.0000	0.9208
X_3	0.9122	0.8513	0.8193	-0.9453	0.9208	1.0000