ORIGINAL PAPER

VARIANCE ESTIMATORS USING NON-PARAMETRIC APPROACH UNDER DIFFERENT RANKED SET SAMPLING SCHEMES

NAEEMA BEGUM[1], MUHAMMAD HANIF[1], USMAN SHAHZAD[1,2*], NASIR ALI[1]

Manuscript received: 12.02.2023; Accepted paper: 07.08.2023;
Published online: 30.09.2023.

Abstract. *Estimation of variance is a commonly discussed topic under simple random sampling (SRS) scheme. The current article deals the issue of variance estimation utilizing supplementary information with the nonparametric approach under different ranked set sampling (RSS) schemes. We propose a class of nonparametric variance estimators utilizing kernel regression [1] with different bandwidths (Plug-in and Cross-validation), under RSS schemes. Simulation study is provided utilizing diverse data sets. The comparison of simulation results has been made between the members of the proposed class with respect to the unbiased variance estimator.*

Keywords*: auxiliary information; nonparametric kernel estimator; ranked set sampling; variance estimation.*

1. INTRODUCTION

Regression analysis (R.A.) is a widely used statistical technique for modelling and examining the connection between the research variables and one or more predictors. It helps to determine how and to what extent the research variable changes as a result of variations in the predictor variables.

The R.A. is predicated on a number of strict assumptions, the two most essential of which are the assumption regarding the error distribution and the established association between the study and predictor variables. With regard to real-world data, these assumptions are actually not always true. By modifying nonparametric regression, which is thought of as an alternate approach, the issue of not always satisfying is resolved.

1.1. NON-PARAMETRIC REGRESSION

Let $(x_i, y_i) \in$ R be given bi-variate data for auxiliary and study random variables. It is well known that the regression model, with $m(x)$ as unknown regression function and $\in_i$ as a random variable of error with zero mean and σ_i^2 variance, is given by

$$y_i = m(X_i) + \in_i , \qquad for\ i = 1, 2, \dots n \tag{1}$$

For the given data, the most common nonparametric method for estimating $m(x)$ is the Kernel estimator or kernel smoother (KS) that is initially proposed by which was first put

[1] Pir Mehr Ali Shah Arid Agriculture University Rawalpindi, Department of Statistics, 46300 Rawalpindi, Pakistan.
[2] International Islamic University, Department of Mathematics and Statistics, 44000 Islamabad, Pakistan.
* *Corresponding Author: usman.stat@yahoo.com.*

out by [2]. The KS estimator is defined, in a fixed neighborhood or area around any x estimation point, as the weighted average value of responses. Mathematically, the KS estimator of $m(x)$ for any x is given by

$$\widehat{m}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} y_i K\left(\frac{x - X_i}{h}\right), \tag{2}$$

where, h is the bandwidth, used to parameterize the weights size. Further, K(.) is the kernel function used to determine kernel weight shape, See [3], for the K(.) the following three conditions are met:

- $\int_{\emptyset_v} K(v)dv = 1$
- $\int_{\emptyset_v} vK(v)dv = 0$
- $\int_{\emptyset_v} v^2K(v)dv \neq 0$

Actually, the KS estimator yields poor results in very uneven x-spaces [4] and [5] solved this problem by introducing a kernel weight as

$$w_i = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

And the KS estimator" named N-W kernel smoother estimator" of $m(x)$ for any x may be

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} \tag{3}$$

Silverman [6] introduced an adaptive N-W kernel weights through the local band width factors λ_i, i.e

$$w_i = \frac{\frac{1}{\lambda_i}K\left(\frac{x - X_i}{\lambda_i h}\right)}{\sum_{i=1}^{n} \frac{1}{\lambda_i}K\left(\frac{x - X_i}{\lambda_i h}\right)}$$

as an extension to the idea of [7]. He used the geometric mean (G.M) in bandwidth factor [1] employed the arithmetic mean (A.M) rather than G.M. The adaptive N-W kernel smoother estimator of $m(x)$ for any x be

$$w_i = \frac{\sum_{i=1}^{n} \frac{y_i}{\lambda_i}K\left(\frac{x - X_i}{\lambda_i h}\right)}{\sum_{i=1}^{n} \frac{1}{\lambda_i}K\left(\frac{x - X_i}{\lambda_i h}\right)} \tag{4}$$

1.2. ESTIMATION OF POPULATION VARIANCE

In our day to day life, variations are present everywhere. It is the nature of law that individuals or no two things are exactly the same. We do not need to emphasize the benefits of difference in all human beings, nature, and creatures. Suffice it to ask: What if there is only

one category of what we eat, drink or see from nature around us? Then what if there was a day without a night despite the dire need of the sun in our lives? How will it live if it rains non-stop, or if drought prevails at all times and places? In order not to keep talking in generalities, the problem is not the principle of disparities or differences within humans, societies, and countries, but rather how to find a full understanding of that variations, how to reduce or increase it or how to find, with the presence of these variations, appropriate estimate for the population parameter that can be used to make decisions and put the right plans. For instance, an agriculturist needs an adequate understanding of the variations in climatic factors especially from place to place (or time to time) to be able to plan on when, how and where to plant his crop. For constant knowledge of the level of variations in people's reactions, a manufacturer needs to reduce or increase the price of his product, or improve the quality of his product. A physician needs a full understanding of variations in the body temperature, degree of human blood pressure and pulse rate for a full prescription.

A population parameter is a function that can be calculated based on the values of one or more population-specific variables. These variables might be the ones that should be approximated. In sample surveys, the sampling design may occasionally be linked to one or more auxiliary variables. These auxiliary variables are frequently employed to enhance designs and to increase estimation accuracy for such unknowable population factors, such as the mean, total, or variance of a research variable. In sampling practice, the estimation of population variance of the research variable is an ongoing topic, and many decisions have been taken to increase estimate accuracy utilizing auxiliary data.

Using supplementary information to increase the effectiveness of the estimates isn't new in sample surveys (see, e.g., [8-10]). What is surprising is the restricted utilization of nonparametric approach (see, e.g., [1, 11-18], disregarding its across the board use in survey practices based on regression estimation. It might intrigue, hence, to discover whether diverse adjustment strategies proposed in the writing for nonparametric regression can improve the performance of the estimates of the population variance. The objective of the paper is committed to this point.

After a brief discussion on nonparametric regression and estimation of population variance, the rest of the paper is structured as follows. Section 2 consists of the adapted estimators. Section 3 consists of the proposed class of nonparametric variance estimators under RSS schemes. Section 4 focuses on presenting the numerical illustration along with its computational results of the simulation study and real-life data. Section 5 covers the most important remarks drawn from the obtained results.

2. MATERIALS AND METHODS

2.1. ADAPTED ESTIMATORS UNDER SRS

According to the law of total variance see [19],

$$var(Y^a) = E[var(Y^a Ix)] + var(\mu_x^a) \tag{6}$$

where, $\mu_x^a = E(Y^a Ix)\ for\ a = 1,2$. Hence the estimates of μ_x^a for $a = (1,2)$, can be incorporated for the estimation of $E(Y^a)$ for $a = (1,2)$.

The parameters μ_x^a for $a = (1,2)$ can be estimated by taking the average over n estimates of $E(yIX_i)$ $i = 1,2,\dots,n$, $a = 1,2$ We propose to estimate the quantities $E(y^a IX = x)$ by utilizing Nadaraya-Watson (NW) and, [1], nonparametric kernel regression

functions. Let K(.) be a kernel function and $h_a > 0$ for $(a = 1, 2)$ be the bandwidth, then these quantities can be estimated using the weighted average.

$$m(x) = \begin{cases} m(x, h_1) = \dfrac{\sum_{i=1}^{N} y_i^1 K\left(\dfrac{x - X_i}{h_i}\right)}{\sum_{i=1}^{N} K\left(\dfrac{x - X_i}{h_i}\right)}, For\ Nadaraya\ and\ Watson\ a = 1 \\ m(x, h_1, \lambda_J) = \dfrac{\sum_{i=1}^{N} \dfrac{Y_i^1}{\lambda_j^1} K\left(\dfrac{x - X_i}{h_i}\right)}{\sum_{i=1}^{N} \dfrac{1}{\lambda_j^1} K\left(\dfrac{x - X_i}{\lambda_i h_i}\right)}, for\ Demir\ and\ Toktamis\ (2010), a = 1 \end{cases} \quad (7)$$

$$m(x) = \begin{cases} m(x, h_2) = \dfrac{\sum_{i=1}^{N} y_i^2 K\left(\dfrac{x - X_i}{h_2}\right)}{\sum_{i=1}^{N} K\left(\dfrac{x - X_i}{h_2}\right)}, For\ Nadaraya\ and\ Watson\ a = 1 \\ m(x, h_2, \lambda_J) = \dfrac{\sum_{i=1}^{N} \dfrac{Y_i^1}{\lambda_j^1} K\left(\dfrac{x - X_i}{h_i}\right)}{\sum_{i=1}^{N} \dfrac{1}{\lambda_j^2} K\left(\dfrac{x - X_i}{\lambda_j h_2}\right)}, for\ Demir\ and\ Toktamis\ (2010), a = 1 \end{cases} \quad (8)$$

In generalized form we can express $m(x)$ as follow

$$m(x) = \begin{cases} m(x, h_a) = \dfrac{\sum_{i=1}^{N} y_i^a K\left(\dfrac{x - X_i}{h_a}\right)}{\sum_{i=1}^{N} K\left(\dfrac{x - X_i}{h_a}\right)}, For\ Nadaraya\ and\ Watson\ a = 1 \\ m(x, h_a, \lambda_J) = \dfrac{\sum_{i=1}^{N} \dfrac{Y_i^a}{\lambda_j^a} K\left(\dfrac{x - X_i}{h_a}\right)}{\sum_{i=1}^{N} \dfrac{1}{\lambda_j^a} K\left(\dfrac{x - X_i}{\lambda_i h_a}\right)}, for\ Demir\ and\ Toktamis\ (2010), a = 1 \end{cases} \quad (9)$$

In case of sample-based study, (9) can be written as

$$\widehat{m}(x) = \begin{cases} \widehat{m}(x, h_a) = \dfrac{\sum_{i=1}^{N} y_i^a K\left(\dfrac{x - X_i}{h_a}\right)}{\sum_{i=1}^{N} K\left(\dfrac{x - X_i}{h_a}\right)}, For\ Nadaraya\ and\ Watson\ a = 1 \\ \widehat{m}(x, h_a, \lambda_J) = \dfrac{\sum_{i=1}^{N} \dfrac{Y_i^a}{\lambda_j^a} K\left(\dfrac{x - X_i}{h_a}\right)}{\sum_{i=1}^{N} \dfrac{1}{\lambda_j^a} K\left(\dfrac{x - X_i}{\lambda_i h_a}\right)}, or\ Demir\ and\ Toktamis\ (2010), a = 1 \end{cases} \quad (10)$$

It is worth mentioning that Gaussian and Epanechnikov are the two most widely used kernels are considered for the purposes of this article. As we know that kernel functions based on the parameter h i.e. bandwidth. Hence, two techniques considered for bandwidth selection namely cross-validation (CV) method, due to [20], and, plug-in method, due to [21]. So, in light of (5), (6) and (10), they propose a class of estimators for the estimation of population variance under simple random sampling, as given below:

$$\hat{\sigma}_{p1}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_2^{Pb}\right) - \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_1^{Pb}\right)\right)^2$$

$$\hat{\sigma}_{p2}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_2^{CV}\right) - \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_1^{CV}\right)\right)^2$$

$$\hat{\sigma}_{p3}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_2^{Pb}, \lambda_{j1}\right) - \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_1^{Pb}, \lambda_{j1}\right)\right)^2$$

$$\hat{\sigma}_{p4}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_2^{Pb}, \lambda_{j2}\right) - \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_1^{Pb}, \lambda_{j2}\right)\right)^2$$

$$\hat{\sigma}_{p5}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_2^{CV}, \lambda_{j1}\right) - \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_1^{CV}, \lambda_{j1}\right)\right)^2$$

$$\hat{\sigma}_{p6}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_2^{CV}, \lambda_{j2}\right) - \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{m}\left(x_i, h_1^{CV}, \lambda_{j2}\right)\right)^2$$

where h^{PB} denotes plug-in bandwidth selector, h^{CV} denotes cross-validated bandwidth selector. Further λ_{j1} , λ_{j2} denotes modified local bandwidth factors based on A.M and G.M, respectively, see [1]. Note that, the proposed estimators $\hat{\sigma}_{p1}^2 - \hat{\sigma}_{p6}^2$ are provided in compact form. So, interested readers may use the full forms of $\widehat{m}(x,.,.)$ and $\widehat{m}(x,.,)$ for $(a = 1, 2)$ available in (10), and get detailed expressions of proposed class.

2.3. PROPOSED VARIANCE ESTIMATORS UNDER RSS SCHEMES WITH SOME BANDWIDTH SELECTORS

In some circumstances, ranked set sampling (RSS), which can significantly increase precision, is a better option than simple random sampling (SRS). It was initially created by [22] to calculate herbage production in agriculture. When measuring a unit directly is expensive or time-consuming but ranking a small group of experimental units is simple and inexpensive, the RSS is preferred. RSS is more accurate than simple random sampling with replacement (SRSWR).

However, no significant work is available regarding non-parametric variance estimation under *RSS*. So we propose a class of non-parametric variance estimators using different bandwidth selection methods and different *RSS* schemes as given below:

$$\hat{\sigma}_{p_j(RSS)}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_{b_j(RSS)}(x) - \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{m}_{a_j(RSS)}(x)\right)^2$$

where

$$\widehat{m}_{a_j(RSS)}(x) = \frac{\sum_{i=1}^{n} y_{[i]} K\left(\frac{x - X_{(i)}}{h_j}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_{(i)}}{h_j}\right)} \tag{11}$$

$$\widehat{m}_{b_j(RSS)}(x) = \frac{\sum_{i=1}^{N} y_{[i]}^2 K\left(\frac{x - X_{(i)}}{h_j}\right)}{\sum_{i=1}^{N} K\left(\frac{x - X_{(i)}}{h_j}\right)} \tag{12}$$

Traditional RSS, $MRSS$, and $ERSS$ are used. The details of these RSS schemes are provided in the next sub-section.

2.4. RANKED SET SAMPLING SCHEMES

2.4.1. Ranked Set Sampling (RSS)

To minimize ranking error in the RSS procedure, the basic step is to choose a set of size m that is typically small, around three or four. The number m refers to the number of sample units assigned to each set. Let's call the study and the auxiliary variables Y and X, respectively. The RSS procedure is summarized in the following five steps:
Step 1: From the population, randomly select m^2 bivariate sample units.
Step 2: As randomly as possible, allocate m^2 selected units into m sets each of size m.
Step 3: Each sample is ranked by one of the variables Y or X. In this case, we assume that the perfect ranking is based on X, whereas the ranking of Y is with possible error.
Step 4: The unit with the smallest rank of X is then measured from the first sample, along with variable Y associated with the smallest rank of X. The variable Y associated with the second smallest rank of X is measured from the second sample of size m. The process is repeated until the Y associated with the highest rank of X is measured from the m^{th} sample.
Step 5: For r cycles, repeat Steps 1–4 until the appropriate sample size, $n = mr$, is achieved for analysis. As an example, we use $SRSWR$ to select a sample of size 36 from a population.

These data are divided into three sets, each of size three, and the process is repeated four times. With RSS methodology, the X values are ranked from smaller to larger in accordance assuming that there is no judgment error. The smallest unit is then chosen from each ordered set, starting with the first ordered set's smallest unit, followed by the second ordered set's smallest unit, and finally the third ordered set's smallest unit. This method yields $n = mr = 12$ observations. Fig. 1 depicts a ranked set sample scheme with set size $m = 3$ and the number of sampling cycles $r = 4$. Despite the fact that 36 sample units were drawn from the population, only the 12 circled units were included in the final sample for quantitative analysis.

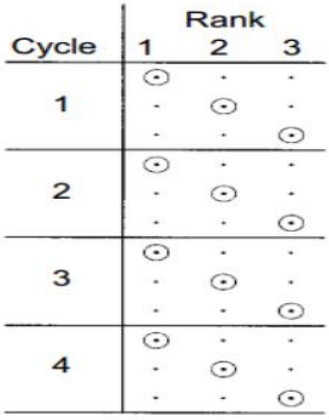

Figure 1. RSS selection procedure

2.4.2. Median Ranked Set Sampling (MRSS)

Muttlak proposed the MRSS method for calculating the population mean. Select m random samples of size m units from the population and use the MRSS technique to rank the units within each sample in relation to an important variable. When the sample size m is odd, choose the median of the sample ((m + 1)/2)th smallest rank) for measurement. When the sample size is even, choose the (m/2)th smallest rank from the first m/2 samples for measurement, and the ((m + 2)/2)th smallest rank from the second m/2 samples. To get mr units of the MRSS data, repeat the operation r times.

According to [22], ranked set sampling (RSS) presupposes flawless ranking, or a lack of ranking errors. However, for the majority of real applications, it is difficult to rank the units accurately. The inaccuracies in ranking the units will result in a loss of precision. When ranking is based on a concomitant variable in this study, the MRSS is employed to estimate the population mean of an important variable. The population mean of the variable of interest is estimated by the regression estimator using an auxiliary variable. The employment of MRSS is more effective, i.e., produces results with lower variance than RSS, for all the scenarios investigated, when one compares the performance of the estimator to that of RSS and regression estimators. For the majority of the scenarios investigated in this study, unless the correlation between the variable of interest and the auxiliary is greater than 90%, the usage of MRSS also yields significantly better results in terms of relative accuracy compared to the regression estimator.

2.4.3. Extreme Ranked Set Sampling (ERSS)

Use the ERSS process to choose n random samples from the population, each of size n units, then visually rank the units within each sample in relation to an interest variable. When the sample size n is even, choose the smallest unit from n/2 samples and the largest unit from the remaining n/2 samples for the actual measurement. When the sample size is odd, choose the smallest unit from (n-1)/2 samples, the largest unit from the other (n-1)/2 samples, and the median of the sample from one sample for the measurement. The procedure may be carried out r times to produce n units of ERSS data. In practice, the $ERSS$ can be performed with fewer errors in ranking the units because all that is required is to find the smallest or largest unit and measure it. The $ERSS$ method is simple to use in the field and will save time when ranking units according to the variable of interest. Furthermore, when compared to RSS, this method will reduce ranking errors and thus increase the efficiency of the $ERSS$.

The following steps are outlined for ERSS:

1. Picks "m" simple random samples with sizes ranging from 1 to m.
2. Sort the elements of each sample, either visually or through another means. Approach that is somewhat cheap yet does not actually measure the desired attribute.
3. Step (3) is carried out once more on an additional m samples of sizes 1, 2,..., m, but this time the minimum ordered observations are measured as opposed to the maximum ordered observations.
4. Measure accurately the maximum ordered observation from the first set, the maximum ordered observation from the second set. The process continues in this way until the maximum ordered observation from the last m-th sample is measured.
5. If more samples are needed, the entire cycle can be performed numerous times.

This variation of RSS preserves some of the balance inherited from the standard RSS in addition to being simpler to run than both the standard RSS and fixed size extreme RSS. As a result, it is anticipated to perform effectively across a larger variety of distributions.

It should stressed here that, even though Φ_a are identified for the study, only $2m$ observations are actually measured and all different ranks are assumed to be obtained with negligible cost and without actual quantification. Hence it is reasonable to compare the sample obtained using this procedure with a simple random sample of size $2m$ and not of size $m(m+1)$.

2.5. BANDWIDTH SELECTION METHODS

The bandwidth determines how close to r two points must be for the assessment of their density to be affected. The estimation is close to the data since a tiny bandwidth only takes the nearest values into account. A smoother estimation results from a large bandwidth's consideration of more points. The bandwidth controls how close two points must be to one another in order for the assessment of their density to be impacted. A small bandwidth only considers the values that are closest, therefore the estimation is close to the data. A wide bandwidth takes into account more points, which leads to a smoother estimation.

It is a well-known fact that bandwidth selection plays a vital role in nonparametric kernel-based methods. We will use the following bandwidths for purposes of the article:

- Plug-in bandwidth of Altman and Leger [21].
- Plug-in bandwidth of Polansky and Baker [23]
- Cross-Validation Bandwidth of Bowman et al. [20].

2.5.1. Proposed Variance Estimators Under RSS

Let $(x_i , yi) \in R$. It is common knowledge that the regression model is given by

$$yi = (x_i) + \in_i , \text{ where } i = 1, 2,..., n \tag{13}$$

and $m(x)$ is an unknown regression function with a zero mean and δ^2 variance.

The Kernel estimator or kernel smoother (KS), which was first presented by [2], is the most used nonparametric technique for estimating m(x) for the provided data. The weighted average value of the responses is what is meant when the KS estimator is defined as the value in a fixed neighborhood or area surrounding any x estimation point. Mathematically, the KS estimator of (x) for any x is given by

$$\hat{m}_h\ (x) = \frac{1}{n_h}\ \sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right) \tag{14}$$

where, h is the bandwidth, used to parameterize the weights size. Further, K (.) is the kernel function used to determine kernel weight shape, See [3], for the K (.) the following three conditions are met:

$$\int \emptyset v\ (v) = 1$$
$$\int \emptyset v\ (v) = 0$$
$$\int \emptyset v\ v\ 2(v) \neq 0.$$

Actually, the Kernel smoothing estimator yields poor results in very uneven x-spaces. [4] and [5] solved this problem by introducing a kernel weight as

$$wi = \frac{k\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n}\left(\frac{x-x_i}{h}\right)}$$

And the Kernel smoothing estimator ‘‘named N-W kernel smoother estimator” of (x). For any x may be

$$\hat{m}_h\ (x)\ = \frac{\sum_{i=1}^{n} y_i k\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right)} \tag{15}$$

Motivated by previous studies, we have proposed variance estimator under ranked set sampling with different band-width selectors

$$\hat{\sigma}^2_{p_j(RSS)} = \frac{1}{n}\sum_{i=1}^{n}\ \hat{m}_{b_j(RSS)}(x;h_j)\ -\ \left(\frac{1}{n}\sum_{i=1}^{n}\hat{m}_{aj(RSS)}\left(x;h_j\right)\right)^2 \text{ For } h_{j=h}$$

h denotes plug-in bandwidth selection estimators

$$\hat{\sigma}^2_{p_1(RSS)} = \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{bj(RSS)}\left(x;h\ \right)\ -\ \left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\hat{m}_{a_j(RSS)}(x;h)\right)^2 \qquad \text{For } h_j = h_{CV}$$

Cv denotes cross validation of [20]

$$\hat{\sigma}^2_{p_2(RSS)} = \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{bj(RSS)}\left(x;h_{CV}\right)\ -\ \left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\hat{m}_{a_j(RSS)}(x;h_{CV})\right)^2 \qquad \text{For } h_{j=}h_{AL}$$

(AL denotes Plug-in bandwidth of Altman and Legger [21]

$$\hat{\sigma}^2_{p_3(RSS)} = \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{bj(RSS)}\left(x;h_{AL}\right)\ -\ \left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\hat{m}_{a_j(RSS)}\left(x;h_{AL}\quad\right)\right)^2 \qquad \text{For } h_j = h_{PB}$$

PB denotes Polansky and Baker plug-in estimates

$$\hat{\sigma}^2_{p_4(RSS)} = \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{bj(RSS)}\left(x;h_{PB}\right)\ -\ \left(\frac{1}{n}\sum_{i=1}^{n}\hat{m}_{a_j(RSS)}(x;h_{PB})\right)^2.$$

Here

$$\hat{m}_{a_j(RSS)}(x) = \frac{\sum_{i=1}^{n} y_{[i]}K\left(\frac{x-X_{(i)}}{h_j}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_{(i)}}{h_j}\right)}$$

$$\widehat{m}_{b_j(RSS)}(x) = \frac{\sum_{i=1}^{N} y_{[i]}^2 K\left(\frac{x - X_{(i)}}{h_j}\right)}{\sum_{i=1}^{N} K\left(\frac{x - X_{(i)}}{h_j}\right)}$$

here $K\left(\frac{x-X_{(i)}}{h_j}\right)$ are kernel function.

For assessing the merits of estimators $\hat{\sigma}_{p1}^2 - \hat{\sigma}_{p4}^2$ we have made an attempt to find the theoretical MSE expressions of the estimators. But we could not obtain these due to mathematical difficulties [1]. So, let we move towards simulation study. As the current article is an initial step for estimation of population variance, under SRSWOR, designed base approach. So, the objective of simulation study is the comparison between proposed estimators on behalf of percentage relative efficiency with respect to unbiased variance estimator, in next section.

3. RESULTS AND DISCUSSION

3.1 RESULTS

To find out the results of our study we simulate data by using Monte Carlo simulation. Our aim is to study efficient non parametric variance estimator under ranked set sampling schemes. We compare the performance estimator and check efficiency. We have used different population to check the behavior of proposed estimators $\hat{\sigma}_{p_1(RSS)}^2$, $\hat{\sigma}_{p_2(RSS)}^2, \hat{\sigma}_{p_3(RSS)}^2, \hat{\sigma}_{p_4(RSS)}^2$ with the existing estimators $\hat{\sigma}_{p_1(SRS)}^2$, $\hat{\sigma}_{p_2(SRS)}^2, \hat{\sigma}_{p_3(SRS)}^2, \hat{\sigma}_{p_4(SRS),\hat{\sigma}_{p_5(SRS)}^2,\hat{\sigma}_{p_6(SRS)}^2}$, by finding their mean square errors and PRE.
Ranking is performed on the auxiliary variable x associated with y. The mean square errors of suggested estimators under ranked set sampling using simulation through R software and compared them with existing mean square errors of estimators under simple random sampling also find their percentage relative efficiency by using the formula:

$$\text{PRE} = \frac{MSE\ (Existing\ Estimator)}{MSE(Proposed\ Estimator)} \times 100$$

We have used following two data sets from various sources to compare the performance of all estimators.

Population 1. Sweden is divided into 284 municipalities for administrative purposes. A municipality is typically made up of a town and its surrounding territory. The size and characteristics of the municipalities vary greatly. We chose a few characteristics to characterize the municipalities in various ways. Official statistics have data on these variables easily available. The generated data set is shown below, and it is utilized in a variety of end-of-chapter exercises to demonstrate key concepts in the book. The data set also allows the reader to conduct his or her own sampling and estimation experiments. The MU284 population refers to the total population of the 284 municipalities. The variables are referred to by their shortened names.
The MU200 population, which consists of the 200 smallest municipalities according to the value of p85, and the MU281 population, which consists of all municipalities other than the

three largest municipalities according to the value of p85, are two smaller populations that we occasionally take into account. This suggests that, with the exception of Stockholm, Goteborg, and Malmo, all municipalities make up the MU281 population. We consider the following variables from MU284 data:

P85: 1985 population (in thousands).

CS82: Number of Conservative seats in municipal council.

Table 1. MSE (Mean Square Error)

MSE		SRS	RSS	MRSS	ERSS
	H	2663304	43719641	49955592	119296352
n=10	C_V	2656520	43361400	49195882	118776853
	A_L	2649111	42757262	48193260	117760772
	P_B	2663740	43654475	49800140	119223802
	H	1592966.6	21671375	49955592	63199777
n=16	C_V	1308204.6	20978133	49195882	62524662
	A_L	932701.6	20202315	48193260	60460253
	P_B	1545854.8	21598603	49800140	63339048
	H	14523.43	24155447	32809156	74492905
n=21	C_V	11613.99	23528216	32438449	73553722
	A_L	11636.10	22830301	31514543	70388780
	P_B	12829.66	24116894	32795057	74638858

Table 2. PRE (Percentage Relative Efficiency)

PRE		SRS	RSS	MRSS	ERSS
	H	100.0000	100.0000	100.0000	100.0000
n=10	PRE.H_C_V	100.2536	100.8262	101.5443	100.4374
	PRE.H_A_L	1005357	102.2508	103.6568	101.3040
	PRE.H_P_B	99.9836	100.1493	100.3122	100.0609
	H	100.0000	100.0000	100.0000	100.0000
n=16	PRE.H_C_V	121.7674	103.3046	102.4834	101.0797
	PRE.H_A_L	170.7906	107.2717	105.4664	104.5311
	PRE.H_P_B	103.0476	100.3369	100.6048	99.7801
	H	100.0000	100.0000	100.0000	100.0000
n=21	PRE.H_C_V	125.0511	102.6639	101.1428	101.2768
	PRE.H_A_L	124.8135	105.8043	104.1080	105.8306
	PRE.H_P_B	113.2020	100.1599	100.0430	99.8044

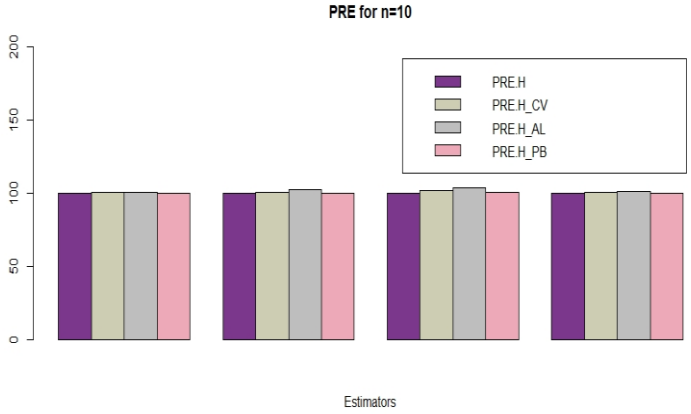

Figure 1. (Percentage Relative Efficiency for n=10)

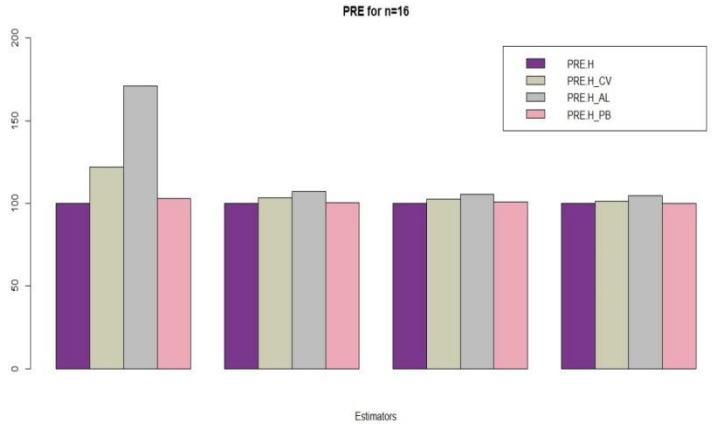

Figure 2. Percentage Relative Efficiency for n=16

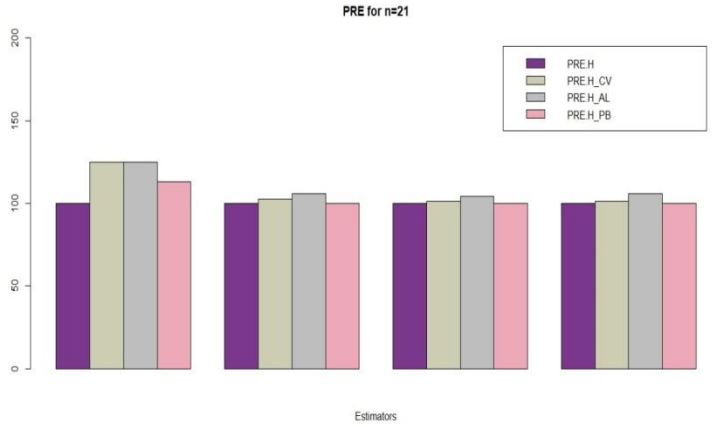

Figure 3. Percentage Relative Efficiency for n=21

Population 2. We consider Murthy data-set. Where X = Data on number of workers and Y = Output for 80 factories in a region

Table 3. MSE (Mean Square Error)

MSE		SRS	RSS	MRSS	ERSS
	H	490668971650	1.393957e+12	1.419615e+12	1.457796e+12
n=9	C_V	411157508005	1.154537e+12	1.232550e+12	1.259629e+12
	A_L	417852604961	1.160714e+12	1.241334e+12	1.265744e+12
	P_B	424900254849	1.173253e+12	1.253376e+12	1.276749e+12
	H	335188519830	1.025337e+12	1.044978e+12	1.086804e+12
n=12	C_V	279809254113	8.845465e+11	8.861653e+11	9.077712e+11
	A_L	284694420046	8.886140e+11	8.916477e+11	9.134391e+11
	P_B	289730540986	8.969139e+11	8.998770e+11	9.233027e+11
	H	249419160603	724395862263	744755302433	1.127410e+12
n=16	C_V	205379942352	618650608413	628055539553	1.067964e+12
	A_L	208719554348	622010499806	632895974214	1.062861e+12
	P_B	212359423313	627660674760	640038443810	1.061909e+12

Table 4. PRE (Percentage Relative Efficiency)

PRE		SRS	RSS	MRSS	ERSS
	PRE.H	100.0000	100.0000	100.0000	100.0000
n=9	PRE.H_C_V	119.3384	120.7374	115.1771	115.7322
	PRE.H _A_L	117.4263	120.0949	114.3621	115.1731
	PRE.H_P_B	115.4786	118.8114	113.2633	114.1804
	PRE.H	100.0000	100.0000	100.0000	100.0000

n=12	PRE.H_C_V	119.7918	115.9167	117.9213	119.7222
	PRE.H_A_L	117.7362	115.3861	117.1963	118.9793
	PRE.H_P_B	115.6897	114.3184	116.1245	117.7083
	PRE.H	100.0000	100.0000	100.0000	100.0000
n=16	PRE.H_C_V	119.3384	120.7374	115.1771	115.7322
	PRE.H_A_L	117.4263	120.0949	114.3621	115.1731
	PRE.H_P_B	115.4786	118.8114	113.2633	114.1804

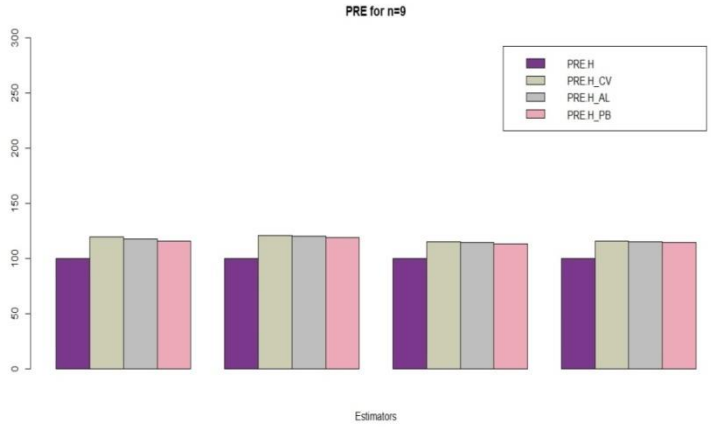

Figure 4. PRE (Percentage Relative Efficiency)

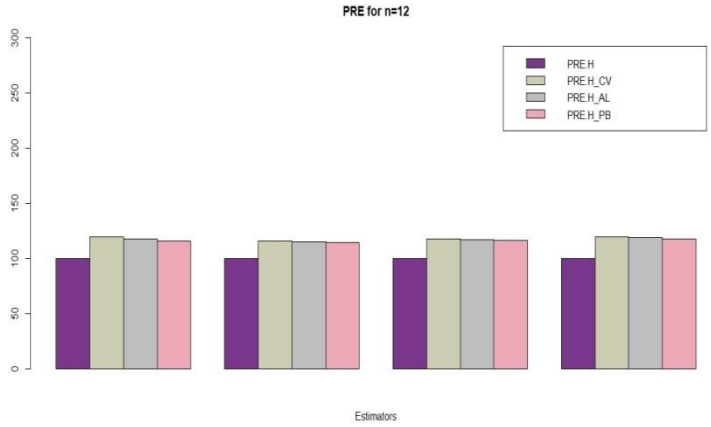

Figure 5. Percentage Relative Efficiency

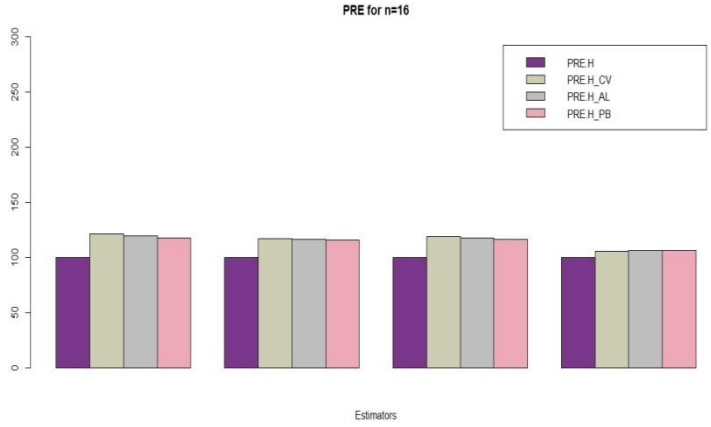

Figure 6. Percentage Relative Efficiency

3.2. DISCUSSION

The following are the most important findings based on the results of MSE and PRE in Tables 1 and 2:

- For all sample sizes, the MSE values associated with C_{v}, A_L, and P_B under the RSS and MRSS sampling schemes are lower than those associated with the H estimator. Furthermore, this is true for both SRS and ERSS sampling schemes except for the P_B in three circumstances under SRS with $n = 10$ and ERSS with $n = 16$ and 21.
- In only one case under SRS with $n = 21$, C_{v} attains the lowest MSE across all estimators.
- With the exception of SRS with $n = 21$, A_L attains the lowest MSE across all estimators for $n = 10, 16$, and 21 under all sampling schemes followed by C_{v}.
- For all sample sizes, the PRE values associated with C_{v}, A_L, and P_B under the RSS and MRSS sampling schemes exceed 100. As a result, C_{v}, A_L, and P_B are more efficient than the H estimator. Furthermore, this is true for both SRS and ERSS sampling techniques except for the P_B in three circumstances under SRS with $n = 10$ and ERSS with $n = 16$ and 21. In other words, the results of PRE can be summarized w.r.t. sampling scheme and sample size as:

Under RSS and MRSS sampling schemes

$$PRE(A_L) > PRE(C_{\mathrm{v}}) > PRE(P_B) > PRE(H), \text{with } n = 10, 16, 21$$

Under SRS sampling scheme

$$PRE(A_L) > PRE(C_{\mathrm{v}}) > PRE(H) > PRE(P_B), \text{with } n = 10$$
$$PRE(A_L) > PRE(C_{\mathrm{v}}) > PRE(P_B) > PRE(H), \text{with } n = 16$$
$$PRE(C_{\mathrm{v}}) > PRE(A_L) > PRE(P_B) > PRE(H), \text{with } n = 21$$

Under ERSS sampling scheme

$$PRE(A_L) > PRE(C_{\mathrm{v}}) > PRE(P_B) > PRE(H), \text{with } n = 10$$
$$PRE(A_L) > PRE(C_{\mathrm{v}}) > PRE(H) > PRE(P_B), \text{with } n = 16, 21$$

The following conclusions are the most crucial ones based on the MSE and PRE results in Tables 3 and 4:

- The MSE values associated with C_{v}, A_L and P_B under the RSS SRS, RSS, MRSS, and ERSS sampling methods are lower for all sample sizes than those associated with the H estimator.
- The PRE values associated with C_{v}, A_L and P_B under the SRS, RSS, MRSS, and ERSS sampling techniques surpass 100 for all sample sizes. As a result, C_{v}, A_L and P_B are more effective than the H estimator.

In accordance with the SRS, RSS, MRSS, and ERSS sample schemes

$$PRE(C_{\mathrm{v}})PRE(A_L) > PRE(P_B) > PRE(H), \text{with } n = 9{,}12 \; and \; 16$$

Steps of simulation study

The steps of simulation study are provided in following points

- Select a RSS of size n from the population.
- Select Cross-validated and Plug-in band widths based on n sampled values of (x; y) with the methods, quoted in Sect. 2.
- Calculate the functions $\widehat{m}(x,.,.)$ and $\widehat{m}(x,.,)$ (say) of Sect. 2, based on n sampled values of (x; y), with their required bandwidth factors i.e λ_{j1} , λ_{j2}
- Find the results of proposed estimators i.e. $\hat{\sigma}_{p1}^2 - \hat{\sigma}_{p4}^2$ with the help of results, obtained in previous steps.
- Repeat all the above steps, 5000 times.

- Finally, evaluate overall MSE as follows $MSE\left(\hat{\sigma}_{pi}^2\right) = \frac{\sum_{j=1}^{5000}\left(\hat{\sigma}_{pi}^2 - \hat{\sigma}^2\right)^2}{5000}$, where $\hat{\sigma}^2 = \frac{\sum_{j=1}^{5000}\left(\hat{\sigma}_{pi}^2\right)^2}{5000}$ and j is the number of replications.
- Through overall MSE results, we find PRE of each estimator w.r.t unbiased variance estimator.
- According to the results of numerical illustration, it is certainly concluded that every proposed variance estimator is better than the existing unbiased variance estimator.

4. CONCLUSIONS

In this paper, we proposed a class of nonparametric estimators of population variance, driven by the law of total variance. The properties of the suggested estimators are provided using various plug-in and cross-validation approaches for bandwidth selection. The theoretical results are then statistically shown using certain real populations, and the outcomes are summarized by MSE and PRE, respectively, in Tables 1 and 2. After careful observing the numerical results of Table 1, we found that the MSE values of bandwidth selection methods namely, $\boldsymbol{C_v}$, $\boldsymbol{A_L}$, and $\boldsymbol{P_B}$ under the RSS and MRSS sampling schemes are lesser than the MSE values of the $\boldsymbol{H}$ estimator for each sample size. This is also true for both SRS and ERSS sampling schemes except for the $\boldsymbol{P_B}$ in three circumstances under SRS with $\boldsymbol{n = 20}$ and ERSS with $\boldsymbol{n = 25}$ and $\boldsymbol{30}$. Further, the bandwidth selection method $\boldsymbol{C_v}$ attains the lowest MSE across all the estimators only under SRS with $\boldsymbol{n = 30}$, whereas, with the exception of SRS with $\boldsymbol{n = 30}$, the bandwidth selection method $\boldsymbol{A_L}$ attains the lowest MSE across all estimators for varying sample sizes such as $\boldsymbol{n = 20, 25}$, and $\boldsymbol{30}$ under all sampling schemes followed by the bandwidth selection method $\boldsymbol{C_v}$. From the results of Table 2, we observed that the PRE values associated with the bandwidth selection methods such as $\boldsymbol{C_v}$, $\boldsymbol{A_L}$, and $\boldsymbol{P_B}$ under the RSS and MRSS sampling schemes exceeds 100 for each sample size. This shows that the bandwidth selection methods such as $\boldsymbol{C_v}$, $\boldsymbol{A_L}$, and $\boldsymbol{P_B}$ are more efficient than the $\boldsymbol{H}$ estimator. Moreover, this is also true for both SRS and ERSS sampling schemes except for the $\boldsymbol{P_B}$ in three circumstances under SRS with sample size $\boldsymbol{n = 20}$ and ERSS with sample sizes $\boldsymbol{n = 25}$ and $\boldsymbol{30}$. After careful observing the numerical results of Table 3, we found that the MSE values of bandwidth selection methods namely, $\boldsymbol{C_v}$, $\boldsymbol{A_L}$, and $\boldsymbol{P_B}$ under the RSS, SRS, ERSS and MRSS sampling schemes are lesser than the MSE values of the $\boldsymbol{H}$ estimator for each sample size. From the results of Table 4, we observed that the PRE values associated with the bandwidth selection methods such as $\boldsymbol{C_v}$, $\boldsymbol{A_L}$, and $\boldsymbol{P_B}$ under the RSS, SRS, ERSS and MRSS sampling schemes exceeds 100 for each sample size. This shows that the bandwidth selection methods such as $\boldsymbol{C_v}$, $\boldsymbol{A_L}$, and $\boldsymbol{P_B}$ are more efficient than the $\boldsymbol{H}$ estimator. As a result, the suggested estimator outperforms the alternatives in every scenario, making it possible to suggest it to survey practitioners for use in real-world situations.

REFERENCES

[1] Demir, S., *Hacettepe Journal Mathematics Statistics*, **39**(3), 429, 2010.
[2] Rosenblatt, M., *Annals Mathematical Statistics*, **27**(3), 832, 1956.
[3] Hardle, W., *Applied nonparametric regression*, Cambridge University Press, 1994.
[4] Nadaraya, E. A., *Theory Probability Applications*, **9**(1), 141, 1964.

[5] Watson, G. S., *Sankhya: The Indian Journal of Statistics A*, **26**(4), 359, 1964.
[6] Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
[7] Abramson, I. S., *Annals Statistics*, **10**(4), 1217, 1982.
[8] Cochran, W.G., *Sampling techniques*, John Wiley and Sons, 1977.
[9] Wolter, K., *Introduction to variance estimation*, Springer Science and Business Media, Berlin, 2007.
[10] Cingi, H., Kadilar, C., *Advances in sampling theory-ratio method of estimation*, Bentham Science Publishers, Sharjah, 2009.
[11] Demir, S., *Scientific Research Essays*, **7**(27), 2409, 2012.
[12] Khulood, H. A., Lutffiah, I. A., *Scientific Research Essays*, **9**(22), 966, 2014.
[13] Hanif, M., *Communications Statistics-Theory Methods*, **44**(9), 1896, 2015.
[14] Hanif, M., *Cogent Mathematics*, **3**(1), 1179247, 2016.
[15] Hanif, M., Shahzadi, S., Shahzad, U., Koyuncu, N., *Sleyman Demirel niversitesi Fen Bilimleri Enstits Dergisi*, **22**(2), 55480106, 2018.
[16] Hanif, M., Shahzad, U., *Journal Statistics Management Systems*, **22**(3), 563, 2019.
[17] Demir, S., *Hacettepe Journal Mathematics Statistics*, **48**(2), 616, 2019.
[18] Ali, T.H., *Communications Statistics-Theory Methods*, **51**(1), 1, 2019.
[19] Spanos, A., *Probability theory and statistical inference: econometric modeling with observational data*, Cambridge University Press, Cambridge, 1999.
[20] Bowman, A., Hall, P., Prvan, T., *Biometrika*, **85**(4), 799, 1998.
[21] Altman, N., Leger, C., *Journal Statistical Planning Inference*, **46**(2), 195, 1995.
[22] McIntyre, G.A., *Australian Journal of Agricultural Research*, **3**(4), 385, 1952.
[23] Polansky, A. M., Baker, E. R., *Journal of Statistical Computation and Simulation*, **65**(1-4), 63, 2000.