Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.13, No.1, 2023

Non-Parametric Estimator for a Finite Population Total Under Stratified Sampling Incorporating a Hybrid of Data Transformation and Reflection Techniques

Nicholas Mugambi[1*], Romanus Otieno Odhiambo[2], Jacob Oketch Okungu[3]

[1,2,3]Department of mathematics Meru University of science and Technology, P O Box 972-60200 Meru, Kenya

*Email Address of the corresponding author: nichmugambi001@gmail.com

ABSTRACT

Survey sampling methods are used in the estimation of population parameters of interest. This field has received increased demand due to the reliable statistics they produce. Information is extracted from the samples and used to make inferences about the population. In this paper, a nonparametric estimator for a finite population total that addresses the problem of boundary bias is proposed. The properties of this estimator were studied in order to determine its accuracy. The estimator was applied to a simulated data and the analysis was done using R statistical package version i386 4.0.3 and the results of the bias confirmed. The performance of the proposed estimator was tested and compared to the design-based Horvitz-Thompson estimator, the model-based approach proposed by Dorfman and the ratio estimator. This was done by studying both the unconditional and conditional properties of the estimators under the linear, quadratic and exponential mean functions. The proposed estimator outperformed other estimators in quadratic and exponential mean functions and therefore can be recommended for estimation and addressing the boundary problem.

***Keywords**: data transformation, data reflection, boundary bias*

1. Introduction

The intensions of surveys are not only in estimating population target parameters, but also in the estimation of subpopulation characteristics. In sample survey, researchers extract information from samples and use such information in making inference about some population quantities such as the mean, proportion or totals. The collection of information can be done either by the use of sampling methods or census. However, census is a tedious and expensive method as it entails complete enumeration of individuals or units contained in a population. Therefore, statisticians rely on the use of sampling methods which involve the selection of a sample from a population of interest and use information obtained from such samples to get estimators of the whole population (Cochran, 1977).

In survey sampling, the estimation of finite population quantities of interest such as the proportions, averages or totals can be done using nonparametric regression method. Nonparametric regression method was introduced early on in the studies by (Nadaraya, 1964) and (Watson, 1964). According to (Dorfman, 1992) nonparametric regression estimators are considered to be more flexible and robust as compared to the estimators based on parametric

regression. In sample survey, auxiliary information is used estimating finite population parameters of interest. However, the use of auxiliary information in estimation of parameters is a key problem in sample surveys. To address this problem, statisticians assume a working super population model to describe the relationship between the auxiliary variable X and the study variable Y (Dorfman, 1992). This super population working model is used in the prediction of the non-sampled part of the population (Sanchez-Borego, 2009).

1.1 Statement of the Problem

In nonparametric, estimation of population totals relies on the use of kernel smoothers since it's an approach for developing a robust estimator. Generally, kernel smoothers suffer the problem of boundary bias. A nonparametric estimator for a finite population total to address this problem based on a hybrid of data transformation and reflection techniques is proposed in this paper.

1.2 Objective.

i. To propose a nonparametric estimator for a finite population total based on a hybrid of data transformation and reflection techniques.
ii. To study the properties of the proposed estimator
iii. To apply the proposed estimator on a simulated data
iv. To compare the performance of the proposed estimator with existing estimators.

2. Summary of Literature

2.1 Nonparametric Regression

The idea of data exploration using nonparametric regression methods has history of introduction. A regression model summarizes the relationship between two variables X and Y by quantifying the contributions of the explanatory variable X to the survey variable Y. There are four main approaches used in estimating finite population totals in sample surveys; model-based approach, design-based approach, model-assisted approach and design-assisted approach.

A model-based approach is applied in this study. In this approach the distribution is a structure existing in the population itself and is unexplored but capable of being modelled. In this forecast approach, the expectations are over all possible realizations of a linear regression stochastic model linking the study variable Y with a set of auxiliary variables X. In the presence of auxiliary information, statisticians assume a working superpopulation model to describe the relationship between the variable of interest and the set of auxiliary variables. We assume that Y is a function of X, hence we have the model

$$Y_i = m(X_i) + \varepsilon_i \quad (1)$$

2.2 Review of Selected Nonparametric Estimators

The idea of nonparametric estimation methods was first introduced by (Nadaraya, 1964) and (Watson, 1964). It was introduced in the estimation of a regression curve using the model

$$Y = m(x_i) + \delta(x_i)e$$

where m(x) is the smoothing function, e is a random error component with a mean of zero and a constant finite variance. The objective of their paper was in estimating the smoothing function m(x). The N-W estimator of the smooth function is given by

$$\hat{m}(x) = \sum_{i\epsilon s} w_i(x) y_i \quad (2)$$

Orwa et al, (2010) proposed a nonparametric regression approach of a finite population total in model-based framework in the case of a stratified sampling. The estimator was based on the modified N-W kernel estimator and it led to relatively small error. Syengo, (2018) considered local polynomial regression under stratified random sampling in the estimation of finite population totals. The population of interest is divided into strata, a simple random sample is selected without replacement from a stratum and the size of the sample should be sufficiently large. The estimates of the study were found to be asymptotically unbiased and consistent. Kim et al adapted the (Breidt and Opsomer, 2000) local polynomial nonparametric regression estimation to two-stage cluster sampling. A probability sample of clusters is drawn from the population of clusters according to a fixed size design and then subsamples of every sampled cluster were obtained. Breidt and Opsomer, (2005) considered a nonparametric design-based regression estimator based on penalized splines. He suggests that they can be used to improve the efficiency of estimators in situations where linear models are not appropriate and are also easy to be incorporated into more complicated models like the additive semiparametric models.

2.2.1 Data Reflection

The basic idea in this method is to reflect the data points at the origin and work with them. It is used in the reduction of bias problems encountered at the boundaries. Lang'at, (2017) studied robust estimation of finite population total in nonparametric regression incorporating data reflection method. The estimator was under model-based framework. The estimator obtained minimized the boundary bias significantly thus it was superior than all other apart from where the ratio estimator dominated in linear model. In their study. Lang'at et al, (2020) explored nonparametric estimation of finite population total under model-based framework. They used kernel smoother in the construction of the estimator. However, this estimator suffers boundary problems which they catered for by modifying it by the use of reflection technique.

2.2.2 Transformation of Data

This technique was introduced in a study by (Wand et al, 1991). Here, one can take a one-to-one function which is continuous and then a regular kernel estimator is used with the transformed data. (Kaarunamuni and Alberts, 2006) applied a locally adaptive transformation method of boundary correction in kernel density estimation. The method was computationally easy and convenient. They found out that the amount of transformation was dependent upon the estimation point. Their estimator depends on the density function applied. Bii et al, (2020) used a modified transformation of data method in estimating finite population mean.

3. Methods

Let $X_1, X_2, \dots, X_N$ be independent and identically distributed random variables with continuous distribution function. Further, let there be a sample of size n and a kernel function k which is symmetric around the origin. Therefore, the standard kernel density estimator is given as;

$$\hat{m}(X_i) = \frac{1}{nh}\sum_{i=1}^{k} k\left(\frac{x-X_i}{h}\right) \quad (3)$$

Where h is the bandwidth and k is a non-negative integrable smoothing kernel.

3.1. Data reflection technique

Let $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ be the set of n observations from the sample. Under reflection of all the points in the boundary, the data increases in number to give the new set of data of the form, $\{(X_1, Y_1), (-X_1, -Y_1), (X_2, Y_2), (-X_2, -Y_2), \dots, (X_n, Y_n), (-X_n, -Y_n)$. Therefore, the kernel estimate obtained from this estimate is of size 2n. The standard kernel estimator for this method is written as

$$\hat{m}_R(x) = \frac{1}{2nh}\sum_{j=1}^{2n} k\left(\frac{x-X_i}{h}\right) \quad x\epsilon\mathbb{R} \quad (4)$$

This can be written as,

$$\hat{m}_R(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{k\left(\frac{x-X_i}{h}\right) + k\left(\frac{x+X_i}{h}\right)\right\} \quad x\geq 0$$

3.2. Transformation of Data Method

The idea behind transformation is based on transforming the original data $X_1, X_2, \dots, X_N$ through a function g to obtain a transformed data given as $g(X_1), g(X_2), \dots, g(X_N)$. Here, g is a positive, continuous and monotonically increasing function. From the standard kernel estimator, the transformed kernel density estimator is of the form,

$$\hat{m}_T(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-g(X_i)}{h}\right) \quad (5)$$

Where h is the bandwidth and k is a symmetric positive kernel function.

3.3. The Proposed Estimator

The proposed hybrid method was obtained by combining data transformation and data reflection techniques to come up with the superior method of estimation. Given the two formulas, $\hat{m}_R(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{k\left(\frac{x-X_i}{h}\right) + k\left(\frac{x+X_i}{h}\right)\right\}$ and $\hat{m}_T(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-g(X_i)}{h}\right)$, we attain the proposed estimator by combining the two to have,

$$\hat{m}_{RT}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left\{k\left(\frac{x-g_1(X_i)}{h}\right) + k\left(\frac{x+g_2(X_i)}{h}\right)\right\} \quad (6)$$

Where h is the bandwidth, k is the kernel function, $g_1 and\ g_2$ are transformations that were determined. For convenience it was assumed that $g_1 = g_2$ for this study. Following the standard formula for estimating finite population totals given as;

$$\hat{T} = \sum_{i\epsilon s} Y_i + \sum_{i\notin s} \hat{m}(x) \quad (7)$$

The proposed estimator becomes;

$$\hat{T}_{RT} = \sum_{i\epsilon s} Y_i + \sum_{i\notin s} \left\{ \frac{1}{nh} \sum_{i=1}^{n} \left[k\left(\frac{x-g_1(X_i)}{h}\right) + k\left(\frac{x+g_2(X_i)}{h}\right) \right] \right\} \quad (8)$$

The assumptions reviewed by (Bii et al, 2020) are used in deriving the bias and variance of the proposed estimator. Assume the transformation $g(x) = x^2 + 2x + 2$ is non-negative continuous and monotonically increasing functions defined on [0,∞). Further, assume that g_i^{-1} exists, $g_i(0)$=2, $g_i'(0)$=2 and that g'' and g''' and are continuous on [0,∞) where $g_i^{(j)}$ denotes the jth derivative of g_i with $g_i^{(0)}$=g_i and g_i^{-1} denoting the inverse function of g_i, i=1,2. Suppose that m^j is the jth derivative of m and that it exists and is continuous on [0,∞), j=0,1,2 with $m^{(0)}$=m. Furthermore, let x =ch where $0 \leq c \leq 1$. Assume the kernel function k is non-negative symmetric function with support [-1,1] such that it satisfies

$$\text{Bias}\left(\hat{T}_{TR}\right) = E\left(\hat{T} - T\right)$$

$$\text{Bias}\left(\hat{T}_{RT}\right) = E\left(\sum_{i=n+1}^{N} \hat{m}_{RT}(x_i) - \sum_{i=n+1}^{N} m(x)\right) \quad (9)$$

The proposed estimator is given as,

$$\hat{m}_{RT}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left\{ K\left(\frac{x-g_1(X_i)}{h}\right) + K\left(\frac{x+g_2(X_i)}{h}\right) \right\}$$

$$\mathrm{E}(\hat{m}_{RT}(x_i)) = \frac{1}{nh} E\left\{ \sum_{i=1}^{n} K\left(\frac{x-g_1(X_i)}{h}\right) + \left(\frac{x+g_2(X_i)}{h}\right) \right\}$$

$$\sum_{i=n+1}^{N} [(\hat{m}_{RT}(x_i))] = \frac{1}{nh} \sum_{i=n+1}^{N} \left\{ E \sum_{i=1}^{n} \left[K\left(\frac{x-g_1(X_i)}{h}\right) + K\left(\frac{x+g_2(X_i)}{h}\right) \right] \right\} \quad (10)$$

Analyzing the first part of the equation

$$= \frac{1}{nh} \sum_{i=n+1}^{N} \left\{ \sum_{i=1}^{n} E\left[K\left(\frac{x-g_1(X_i)}{h}\right) \right] \right\}$$

$$= \frac{N-n}{nh} \int_0^{\infty} K\left(\frac{x-g_1(X_i)}{h}\right) m(X_i) dX$$

$$= \frac{N-n}{n} \int_{-1}^{c} K(t) \frac{m(g_1^{-1}(c-t)h)}{g_1'(g_1^{-1}(c-t)h)} \mathrm{dt} \quad (11)$$

Using Taylor series expansion of order 2 under condition t=c

$$= \frac{N-n}{n} \int_{-1}^{c} \left\{ \frac{m(g_1^{-1}(0))}{g_1'(g_1^{-1}(0))} - (t-c)h \frac{g_1'(g_1^{-1}(0)) m'(g_1^{-1}(0)) - g_1''(g_1^{-1}(0)) m(g_1^{-1}(0))}{[g_1'(g_1^{-1}(0))]^3} \right.$$
$$+ \frac{h^2}{2}(t-c)^2 \left[\frac{g_1'(g_1^{-1}(0)) m''(g_1^{-1}(0)) - g_1'''(g_1^{-1}(0)) m(g_1^{-1}(0))}{[g_1'(g_1^{-1}(0))]^4} \right.$$
$$\left. \left. - \frac{3g_1''(g_1^{-1}(0)) \{ g_1'(g_1^{-1}(0)) m'(g_1^{-1}(0)) - g_1''(g_1^{-1}(0)) m(g_1^{-1}(0)) \}}{[g_1'(g_1^{-1}(0))]^5} \right] \right\} dt + o(h^2)$$

(12)

Using the assumptions $g^{-1}(0) = 0 \; and \; g'(0) = 2$ the equation reduces to

$$= \frac{N-n}{n}\left\{m(0)\int_{-1}^{c}K(t)dt - 2h\int_{-1}^{c}(t-c)K(t)dt[m'(0) - g_1''(0)m(0)] + \frac{2h^2}{2}\int_{-1}^{c}(t-c)^2K(t)dt\{m''(0) - g_1'''(0)m(0) - 3g_1''(0)[2m'(0) - g_1''(0)m(0)]\}\right\} + o(h^2) \tag{13}$$

For the second part of the equation we have,

$$= \frac{1}{nh}\sum_{i=n+1}^{N}\left\{\sum_{i=1}^{n}E\left[K\left(\frac{x+g_2(X_i)}{h}\right)\right]\right\}$$

$$= \frac{N-n}{nh}\int_{0}^{\infty}K\left(\frac{x+g_2(X_i)}{h}\right)m(X_i)dX$$

Using change of variables technique, we have,

$$= \frac{N-n}{n}\int_{c}^{1}K(t)\frac{m(g_2^{-1}(t-c)h)}{g_2'(g_2^{-1}(t-c)h)}\mathrm{dt} \tag{14}$$

Under Taylor series expansion of order 2 at t=c the equation becomes

$$E(\hat{m}_{RT}(x_i)) = \frac{N-n}{n}\left\{m(0) + 2h\left[\int_{c}^{1}(t-c)K(t)dt\{m'(0) - g_1''(0)m(0)\}\right] - 2h\left[\int_{-1}^{c}(t-c)K(t)dt\{m'(0) - g_2''(0)m(0)\}\right] + \frac{2h^2}{2}\left\{\int_{c}^{1}(t-c)^2K(t)dt[m''(0) - g_1'''(0)m(0) - 3g_1''(0)\{2m'(0) - g_1''(0)m(0)\}]\right\} + \frac{2h^2}{2}\left\{\int_{-1}^{c}(t-c)^2K(t)dt[m''(0) - g_2'''(0)m(0) - 3g_2''(0)\{2m'(0) - g_2''(0)m(0)\}]\right\}\right\} + o(h^2) \tag{15}$$

$$= m(x) + 2h\left\{2m'(0)\int_{c}^{1}(t-c)K(t)dt - g_1''(0)m(0)\int_{c}^{1}(t-c)K(t)dt - g_2''(0)m(0)\left(c + \int_{c}^{1}(t-c)K(t)dt\right)\right\} + \frac{2h^2}{2}\left\{-c^2m''(0) + m''(0)\int_{-1}^{1}(t-c)^2K(t)dt - [g_1'''(0)m(0) + 3g_1''(0)\{2m'(0) - g_1''(0)m(0)\}]\int_{c}^{1}(t-c)^2K(t)dt - [g_2'''(0)m(0) + 3g_2''(0)\{2m'(0) - g_2''(0)m(0)\}]\int_{-1}^{c}(t-c)^2K(t)dt\right\} + o(h^2) \tag{16}$$

Thus, the bias is given as,

$$E[\hat{m}_{RT}(x_i)] - m(x)$$

$$= \frac{N-n}{n}\left\{2h\left[2m'(0)\int_{c}^{1}(t-c)K(t)dt - g_1''(0)m(0)\int_{c}^{1}(t-c)K(t)dt \; - g_2''(0)m(0)\left(c + \int_{c}^{1}(t-c)K(t)dt\right)\right] + \frac{2h^2}{2}\left[-c^2m''(0) + m''(0)\int_{-1}^{1}(t-c)^2K(t)dt - [g_1'''(0)m(0) + 3g_1''(0)\{2m'(0) - g_1''(0)m(0)\}]\int_{c}^{1}(t-c)^2K(t)dt - [g_2'''(0)m(0) + 3g_2''(0)\{2m'(0) - g_2''(0)m(0)\}]\int_{-1}^{c}(t-c)^2K(t)dt\right]\right\} + o(h^2) \tag{17}$$

The estimator is asymptotically unbiased. As $n \to \infty \; and \; h \to 0$ the bias of the estimator tends to zero.

3.4. Variance of the Proposed Estimator.

The variance of the proposed estimator is given as

$$var(T) = E[T]^2 - [E(T)]^2 \tag{18}$$

$$var[\textstyle\sum_{i=n+1}^{N}(\widehat{m}_{RT})] = \frac{(N-n)^2}{nh^2}\left\{var\left[K\left(\frac{x-g_1(X_i)}{h}\right) + K\left(\frac{x+g_2(X_i)}{h}\right)\right]\right\} \tag{19}$$

$$= \frac{(N-n)^2}{nh^2}\left\{E\left[K\left(\frac{x-g_1(X_i)}{h}\right) + K\left(\frac{x+g_2(X_i)}{h}\right)\right]^2 - \left[E\left(K\left(\frac{x-g_1(X_i)}{h}\right) + K\left(\frac{x+g_2(X_I)}{h}\right)\right)\right]^2\right\} \tag{20}$$

We let,

$$A = \frac{(N-n)^2}{nh^2}\left\{E\left[K\left(\frac{x-g_1(X_i)}{h}\right) + K\left(\frac{x+g_2(X_i)}{h}\right)\right]^2\right\}$$

$$= \frac{(N-n)^2}{nh^2}\left\{\int_0^\infty K\left(\frac{x-g_1(X_i)}{h}\right)^2 m(X)dX + \int_0^\infty K\left(\frac{x+g_2(X_i)}{h}\right)^2 m(X)dX + 2\int_0^\infty K\left(\frac{x-g_1(X_i)}{h}\right)K\left(\frac{x+g_2(X_i)}{h}\right)m(X)dX\right\} \tag{21}$$

Using the change of variable technique, by letting $X_i = u$, we have,

$$= \frac{(N-n)^2}{nh^2}\left\{\int_0^\infty K\left(\frac{x-g_1(u)}{h}\right)^2 m(u)du + \int_0^\infty K\left(\frac{x+g_2(u)}{h}\right)^2 m(u)du + 2\int_0^\infty K\left(\frac{x-g_1(u)}{h}\right)K\left(\frac{x+g_2(u)}{h}\right)m(u)du\right\} \tag{22}$$

$$= \frac{(N-n)^2}{nh^2}\left[h\int_{-1}^{c} K^2(t)\frac{m(g_1^{-1}((c-t)h)}{g_1'(g_1^{-1}((c-t)h))}dt + h\int_{c}^{1} K^2(t)\frac{m(g_2^{-1}((t-c)h)}{g_2'(g_2^{-1}((t-c)h))}dt\right] \tag{23}$$

$$= \frac{(N-n)^2 m(0)}{nh}\int_{-1}^{1} K(t)^2 dt + o\left(\frac{1}{nh}\right) \tag{24}$$

By the continuity property of $g_1'' \; and \; g_2''$ and by Taylor expansion of order two on g_1 and g_2, we have,

$$g_1\big((c-t)h\big) = g_1(0) + (t-c)(-h)g_1'(0) + O(h^2)$$

$$= 2 + 2(c-t)h + O(h^2) \tag{25}$$

And,

$$g_2\big((c-t)h\big) = g_2(0) + (t-c)(-h)g_2'(0) + O(h^2)$$

$$= 2 + 2(c-t)h + O(h^2) \tag{26}$$

Since $g_i(0) = 2$ and $g_i'(0) = 2$, i=1,2 using the two equations above and by the change of variables

$$t = \frac{x-g_1(X_i)}{h}$$

$$X_i = g_1^{-1}(x - ht)$$

$$A_2 = \frac{2(N-n)^2}{nh^2}\left\{\int K\left(\frac{x+g_1(X_i)}{h}\right)K\left(\frac{x-g_2(X_i)}{h}\right)m(X_i)dX\right\} \quad (27)$$

$$= \frac{2(N-n)^2}{nh}\left\{\int_{-1}^{c} K(t)K\left(\frac{x-g_2\left(g_1^{-1}(ht-x)\right)}{h}\right)m\left(g_1^{-1}(ht-x)\right)dt\right\} \quad (28)$$

$$= \frac{2(N-n)^2}{nh}\left\{\int_{-1}^{c} K(t)K\left(\frac{x-(2+2(t-c)h+O(h^2))}{h}\right)m(g_1^{-1}(ht-x))dt\right\} \quad (29)$$

$$= \frac{2(N-n)^2}{nh}\int_{-1}^{c} K(t)K(-2+(3c-2t+O(h)))(m(0)+O(h))dt \quad (30)$$

$$= \frac{2(N-n)^2}{nh}\int_{-1}^{c} K(t)K(-2+(3c-2t))dt + o\left(\frac{1}{nh}\right) \quad (31)$$

$$B = \frac{1}{nh^2}\left\{E\left[K\left(\frac{x+g_1(x_i)}{h}\right)+K\left(\frac{x-g_2(x_i)}{h}\right)\right]\right\}^2 \quad (32)$$

$$= o\left(\frac{1}{nh}\right) \quad (33)$$

Therefore, by combining the equations above, we have

$$Var(\hat{m}_{RT}(x)) = \frac{(N-n)^2}{nh}\left\{\int_{-1}^{1} K(t)^2dt + 2\int_{-1}^{c} K(t)K(-2+(3c-2t))dt\right\} + o\left(\frac{1}{nh}\right) \quad (34)$$

The variance of $\hat{m}(x)$ decreases in nh as $n \to \infty$ and the bandwidth h$\to$ 0. This implies that the variance of the estimator converges to zero hence its statistically consistent.

3.5. Mean Squared Error

The mean squared error brings together the variance of the estimator and the square of the bias term of the estimator.

That is,

MSE$(\hat{T}) = E(\hat{T} - T)^2$ or this can be given as

MSE$(\hat{T}) = Var(\hat{T}) + (Bias)^2$

$$= \frac{(N-n)^2}{nh}\left\{\int_{-1}^{1} K(t)^2dt + 2\int_{-1}^{c} K(t)K(-2+(3c-2t))dt\right\} + \left[\frac{N-n}{n}\left\{2h\left[2m'(0)\int_{c}^{1}(t-c)K(t)dt - g_1''(0)m(0)\int_{c}^{1}(t-c)K(t)dt - g_2''(0)m(0)\left(c+\int_{c}^{1}(t-c)K(t)dt\right)\right] + \frac{2h^2}{2}\left[-c^2m''(0) + m''(0)\int_{-1}^{1}(t-c)^2K(t)dt - [g_1'''(0)m(0)+3g_1''(0)\{2m'(0)-g_1''(0)m(0)\}]\int_{c}^{1}(t-c)^2K(t)dt - [g_2'''(0)m(0)+3g_2''(0)\{2m'(0)-g_2''(0)m(0)\}]\int_{-1}^{c}(t-c)^2K(t)dt\right]\right\}^2 + o\left(\frac{1}{nh}\right) \quad (35)$$

Since the estimator is asymptotically unbiased and its variance converges to zero, the mean squared error also converges to zero as sample size increase.

4. Simulation Study

Simulation was done using R statistical software version i386 4.0 using three of the theoretical data variables employed by (Breidt and Opsomer, 2000).

The linear model was used to simulate the first data set. The model is given as

$$Y_i = 1 + 2(x_i - 0.5) + e_i \quad (36)$$

The second data set was obtained through simulation by the use of a quadratic model given as

$$Y_i = 1 + 2(x_i - 0.5)^2 + e_i \quad (37)$$

The third data set was simulated by the use of an exponential model given as

$$Y_i = \exp(-8x_i) + e_i \quad (38)$$

The random variable X is generated as independent and identically distributed U(0,1) and the error component is a standard normal. In all the three data variables, a population of size 1000 was simulated and samples of size 300 are selected from each population and the estimates of the population total and the mean squared error computed.

4.1 Unconditional Properties of the Estimator

4.1.1 Unconditional Biases

The biases of our estimator, the estimator proposed by (Dorfman, 1992), the Horvitz-Thompson estimator and the ratio estimator are computed as $(\hat{T}_{TR} - Y), (\hat{T}_{NW} - Y),$

$(\hat{T}_{HT} - Y)$ *and* $(\hat{T}_R - Y)$ respectively.

Table 1: unconditional Bias of the estimators

MODEL	$\hat{T}_{RT}$	$\hat{T}_{NW}$	$\hat{T}_{HT}$	$\hat{T}_R$
Linear	*212.1953*	*935.7327*	*-16.70931*	*-16.08618*
Quadratic	*12.20103*	*568.9697*	*-30.06625*	*-31.18455*
Exponential	*-2.273007*	*-57.98402*	*-12.75688*	*-5.988498*

From Table 1, some of the values of the biases are negative and others are positive which indicate either underestimation or overestimation. For the linear function, the ratio estimator has the lowest bias followed by the Horvitz-Thompson estimator and the proposed estimator is the third. In quadratic model, the proposed estimator performs the best. In exponential model, the proposed estimator has the lowest bias which indicates that it's the best.

4.1.2 Mean Squared Error

The measures of the mean squared errors were computed for the four data sets and then compared.

$$MSE = \frac{\sum_{i=1}^{300}(\hat{T}_i - T)^2}{300} \quad (39)$$

Table 2 unconditional MSE

MODEL	T_{RT}	$\hat{T}_{NW}$	$\hat{T}_{HT}$	$\hat{T}_{R}$
Linear	*150.0895*	*2918.652*	*0.9306704*	*0.8625511*
Quadratic	*0.4962173*	*1079.088*	*3.013264*	*3.241588*
Exponential	*0.01722187*	*11.20716*	*0.5424597*	*0.1195404*

From Table 2, the ratio estimator has the least MSE followed by the Horvitz-Thompson estimator under the linear function. For the quadratic function, the proposed estimator performed the best followed by the Horvitz-Thompson estimator. For the exponential, the proposed estimator outperformed the other three models.

4.2 Conditional Properties of the Estimator

Here, the samples selected are grouped into groups of size 20 therefore we have 15 groups. The grand mean for each group is computed as

$$\bar{\bar{X}} = \frac{1}{15}\sum_{i=1}^{20} \bar{x}_i \tag{40}$$

The mean estimator is also computed as

$$\hat{\bar{T}}_{TR} = \frac{1}{15}\sum_{i=1}^{20} \hat{T}_{TR\,i} \tag{41}$$

The conditional bias for each group was then computed as $(\hat{\bar{T}}_{TR} - \bar{Y})$

The figures 4.1, 4.2 and 4.3 below show the trend of the conditional bias for each estimator under the three mean functions.

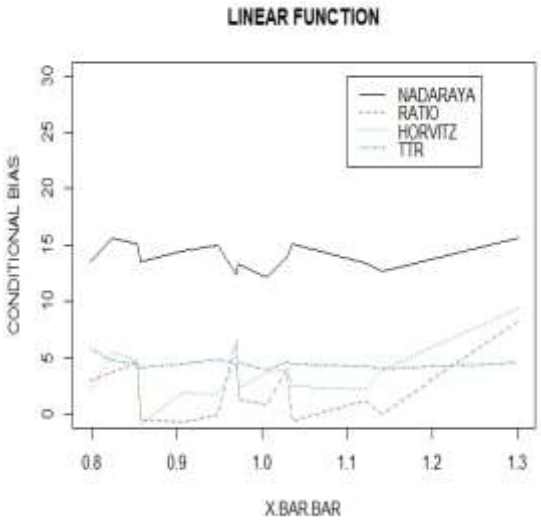

Figure 1 conditional bias linear function

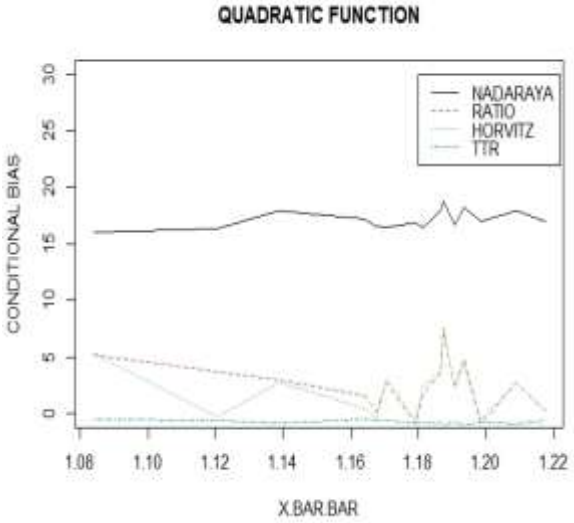

Figure 2 conditional bias quadratic function

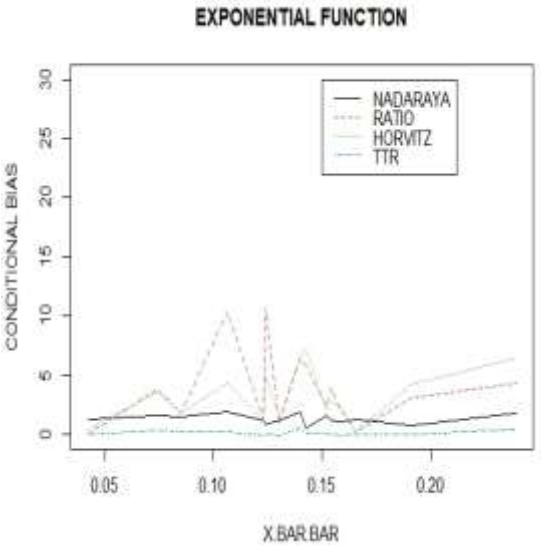

Figure 3 conditional bias for exponential function

From figure 1, where the linear mean function was applied, the ratio estimator gave the best results. This is attributed to the fact that the ratio estimator is the Best Linear Unbiased Estimator (BLUE) thus it cannot be outperformed by any other estimator. It can be observed from the graph that the biases of the estimators are minimal

Figure 2 where a quadratic mean function was applied, the proposed estimator gave the best estimates followed by the Horvitz-Thompson estimator, ratio estimator and the Nadaraya-Watson estimator performed poorly. It can also be observed from the graph that the bias between the estimators is large on the left but reduces towards the right as the mean increases.

From figure 3 the exponential mean function was applied, the proposed estimator gave better estimates of the population total followed by the Nadaraya-Watson estimator, the Horvitz-Thompson estimator and the ratio estimator performed poorly. It can be observed from the graph that the biases are minimal throughout the graph.

5. Conclusions and Recommendations

In this paper, we developed an estimator for finite population total based on a composite of data transformation and data reflection techniques which addressed the problem of boundary bias effectively as shown from the biases in Table 4.1. The proposed estimator was found to perform quite well under the quadratic and exponential models where it produced low biases as compared to the Ratio estimator, Horvitz-Thompson and the Nadaraya-Watson estimator. However, the ratio estimator was the best under linear models since it's the Best Linear Unbiased Estimator (BLUE). Our estimator has the least mean squared error over the two models.

5.1 Recommendations

The proposed nonparametric estimator for a finite population total was developed and it performed better than (Dorfman, 1992) estimator and therefore can be recommended for estimation and addressing the boundary problem. In this paper estimation is carried out using stratified sampling, therefore estimation using cluster sampling is recommended to compare the performance of the estimator. Also, an improvement of the estimator in order to work on all the theoretical data variables are recommended.

ACKNOWLEDGEMENT

Much appreciation goes to all who guided us in one way or the other throughout the period of study. We are greatly indebted to staff of the mathematics department in MUST.

REFERENCES

Bii, N. K., Onyango, C. O., & Odhiambo, J. (2020). *Estimating a finite population mean using transformed data in presence of random nonresponse*. International Journal of Mathematics and Mathematical Sciences, 2020.

Breidt, F. J., & Opsomer, J. D. (2000). *Local polynomial regression estimators in survey sampling*. Annals of statistics, 1026-1053

Breidt, F. J., Claeskens, G., & Opsomer, J. D. (2005). *Model-assisted estimation for complex surveys using penalised splines*. Biometrika, 92(4), 831-846.

Cheruiyot, L. R. (2020). *Exploring Data-Reflection Technique in Nonparametric Regression Estimation of Finite Population* Total: An Empirical Study. American Journal of Theoretical and Applied Statistics, 9(4), 101-105.

Cochran, W.G. (1977), *Sampling techniques, Third edition,* New York: John Wiley and Sons.

Dorfman, A. H. (1992). *Nonparametric regression for estimating totals in finite populations. In Proceedings of the Section on Survey Research Methods* (pp. 622-625). American Statistical Association Alexandria, VA.

Horvitz, D. G., & Thompson, D. J. (1952). *A generalization of sampling without replacement from a finite universe.* Journal of the American Statistical Association, 47(260), 663-685. https://doi.org/10.1080/01621459.1952.10483446.

Karunamuni, R. J., & Alberts, T. (2006). *A locally adaptive transformation method of boundary correction in kernel density estimation.* Journal of Statistical Planning and Inference, 136(9), 2936-2960.

Kim, J. Y., Breidt, F. J., & Opsomer, J. D. (2003). *Nonparametric regression estimation of finite population totals under two-stage sampling.* preprint.

Lang'at, R. C. (2017). *Robust Estimation of Finite Population Total Incorporating Data-Reflection Technique in Nonparametric Regression* (Doctoral dissertation, COPAS, JKUAT).

Nadaraya, E. A. (1964). *On estimating regression*. Theory of Probability & Its Applications, 9(1), 141-142.

Orwa, G. O., Otieno, R. O., & Mwita, P. N. (2010). *Nonparametric mixed ratio estimator for a finite population total in stratified sampling*.

Rueda, M., & Sánchez-Borrego, I. R. (2009). *A predictive estimator of finite population mean using nonparametric regression.* Computational Statistics, 24(1), 1-14.

Syengo, C. K. (2018). *Local Polynomial Regression Estimator of the Finite Population Total under Stratified Random Sampling: A Model-Based Approach* (Doctoral dissertation, JKUAT-PAUSTI)

Wand, M.P., Marron, J.S. and Ruppert, D. (1991). *Transformations in Density Estimation (with discussion).* Journal of the American Statistical Association, 86, 343-361.

Watson, G. S. (1964). *Smooth regression analysis.* Sankhyā: The Indian Journal of Statistics, Series A, 359-372.