ESTIMATION OF FINITE POPULATION TOTALS BASED ON A ROBUST NONPARAMETRIC FEEDFORWARD BACKPROPAGATION NEURAL NETWORK

FESTUS ANZETSE WERE

JOMO KENYATTA UNIVERSITY

OF

AGRICULTURE AND TECHNOLOGY

2023

Estimation of Finite Population Totals Based on a Robust Nonparametric Feedforward Backpropagation Neural Network

Festus Anzetse Were

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Applied Statistics of the Jomo Kenyatta University of Agriculture and Technology

2023

DECLARATION

This Thesis is my original work and has not been presented for a degree in ay other University.

Signature.. Date................................

Festus Anzetse Were

This Thesis has been submitted for examination with our approval as University Supervisors.

Signature.. Date................................

Prof. Romanus Odhiambo Otieno

JKUAT, Kenya

Signature.. Date................................

Prof. George Otieno Orwa

JKUAT, Kenya

DEDICATION

To my parents Mr. and Mrs. Were, my brothers and sisters. Also, I dedicate it my wife and daughter Nadiah Pendo.

ACKNOWLEDGEMENT

I thank the almighty God for granting me the knowledge to start, continue and finalize this study. My sincere gratitude goes to my colleagues for the company and readiness to read this material and their corrections. I can't overlook the love and support from my family and friends. Special thanks to my supervisor Prof. George Otieno Orwa for his presence since the start of this work until where it has reached. I highly appreciate the moral and academic support he is gave me. I do also thank my supervisor Prof. Romanus Odhiambo Otieno for the scholarly and professional assistance he offered and his presence in correcting my work. I also express my sincere gratitude to my colleague Macdonald Obudho for his support during this journey and also the staff of Statistics and Actuarial Sciences Department for the friendly guidance they gave me in this study. May God bless you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

AMISE	Asymptotic Mean Integrated Square Error
ANN	Artificial Neural Network
GAM	Generalized Additive Models
GCV	Generalized Cross Validation
FFNN	Feed Foward Neural Network
i	Sample element
j	Non-sample element
LP	Local Polynomial
MAE	Mean Absolute Error
MARS	Multivariate Adaptive Regression Spline
MAPE	Mean Absolute percentage Error
MLE	Maximum Likelihood Estimator
MSE	Mean Squared Error
MLP	Multilayer Perceptron Neural Network
MISE	Mean Integrated Square Error
NN	Neural Network
RME	Root Mean Error
RAB	Relative Absolute Bias
RAB	Relative Bias
RRMSE	Relative Root Mean Squared Error
SRSWOR	Simple Random Sampling Without Replacement
SSE	Sum of Squared Errors
UN	United Nations
U	Finite population
$\hat{\mathrm{T}}_{\mathrm{NN}}$	Feedforward neural network Estimator
$\hat{\mathrm{T}}_{\mathrm{MARS}}$	Multivariate Adaptive Regression Spline Estimator

$\hat{\mathrm{T}}_{\mathrm{LP}}$	Local polynomial Estimator
$\hat{\mathrm{T}}_{\mathrm{GAM}}$	Generalized Additive Estimator
$\hat{\mathrm{m}}()$	Estimator of the mean function
$\mathrm{m}()$	The mean function

ABSTRACT

Estimation procedure of Population Parameters in Model Based framework has employed Nonparametric techniques widely. This has become more interesting when complete Auxiliary information is available allowing use of more flexible methods in predicting the value taken by the survey variable in nonsampled units ensuring more efficient Estimators of Finite Population Totals are build. In this context, estimators such as Local Polynomial and Kernel Smoothers have dominantly been used and shown to provides good estimators for Finite Population Total in low dimension. Even in these scenarios however, bias at boundary points presents a big problem when using these estimators in estimating Finite Population Parameters. The problem worsens as the dimension of the regressors vectors increases. This leads to sparseness of regressors values in the design space making these methods unfeasible due to the decrease in the fastest achievable rates of convergence of the regression function estimator towards the target curve. To address this challenges, this study considers estimation of Finite Population Totals in high dimension using a Feedforward Backpropagation Neural Network. The technique of Neural Network ensures Robust Estimator in high dimensions and reduces estimation bias with marginal increase in variance. The estimators properties are developed, and a comparison with existing estimators such as Generalized Additive Models, Multivariate Adaptive Regression Spline and Local Polynomial was conducted to evaluate the estimators performance using simulated data and data acquired from the United Nations Development Programme 2020. When certain conditions are met, the estimator was found to have an asymptotic Mean Square Error and asymptotically consistent. Simulation results showed that, the Feedforward Backpropagation Neural Network estimator is efficient and outperformed the existing estimators in estimating Finite Population Totals as it had smaller values of biases, and mean square errors compared to other Estimators. The estimation approach performs well in an example using data from a United Nations Development Programme 2020 on the study of Human Development Index against other factors. The theoretical and practical results imply that the Feedforward Backpropagation Neural Network Estimator is highly recommended for Survey Sampling in the Estimation of Finite Population Totals.

CHAPTER ONE
INTRODUCTION

1.1 Background of the Study

Surveys are conducted at Local, National and International Levels to gather information and aid Public and Private sectors in effective policy making, (Hansen et al., 1987). The objective of any survey is usually to obtain summary Statistics for a Finite Population or for specific subgroups at the same time reducing the time and cost of collecting data. Extrapolation does not give accurate information in Surveys since the Sample is a subset of an entire Population and therefore does not have information on Units that are not represented in the selected Sample. Auxiliary Information that is correlated to the characteristic under study has been very effective in predicting the information in the unobserved units of the Population under study.

For the purposes of this study, suppose there is a Finite Population of N distinct and identifiable units; $U = \{1, 2. \dots, N\}$. Let each Population Unit have the variable of interest as Y. It is assumed that there exist an Auxiliary Variable $X \in \mathbb{R}^d$, closely associated with Y, which is known for the entire Population ($i.e$ $X_1, X_2 \dots, X_N$) are known $\forall$ $Y_i, i = 1, 2, \dots, N$. Often the researchers are faced with the problem of estimating a function of the Population, (i.e a function of $Y's$), such as the Population Total;

$$T = \sum_{i=1}^{N} Y_i \tag{1.1}$$

or the Population Mean $\hat{Y}$ or the Population Distribution Functions

$$F(y) = \frac{1}{N} \sum_{i=1}^{N} I(Y_i \leq y) \tag{1.2}$$

The use of Distribution Functions in Survey Sampling, see (Chambers et al., 1992) .

In estimating the Population Total T for instance, a Sample S is usually taken so that the pair $(x_{i,j}, y_i), i = 1, 2, \ldots, n$ and $j = 1, 2, 3, \ldots, d$ is obtained from the variable X and corresponding variable Y. It may then be used in the Design Stage, Estimation Stage or both, (Hadayat and Silha, 1991). In the presence of such auxiliary variables, a researcher can use a Superpopulation Model at the estimation stage for inference, (Chambers and Dunstan, 1986; Wang and Dorfman, 1996). Thus, estimators are sought to have desirable properties like Asymptotic Design Unbiased, Consistency irrespective of whether the working model is correctly specified or not and to be particular the efficiency of the model.

However, all these techniques refer to simple statistical models for the underlying relationships between the Survey and Auxiliary variable (Linear Regression Models). (Hansen et al., 1983), points out through empirical study that, under the Parametric Superpopulation, Misspecification of the Model can lead to serious errors in an inference. To solve this problem, Nonparametric Regression involving Robust Estimators in Finite Population Sampling has been proposed by (Dorfman, 1992a; Otieno and Mwalili, 2000; Breidt and Opsomer, 2000).

The reason behind the Nonparametric Approach in this study is that a regression curve obtained in this way has four main purposes detailed by (Härdle and Linton, 1994). It provides a versatile method of exploring the general relationship between two variable; secondly, it enables one to make prediction of observations without any reference to fixed Parametric Model; thirdly, it is a tool for finding spurious observations by studying influence of isolated points and lastly it is a flexible method for interpolating between adjacent values of Auxiliary variable.

A major problem that is usually encountered when using Nonparametric Kernel based Regression Estimators over a Finite Interval such as the estimation of Finite Population Quantities is the bias at the boundary points, (Chambers et al., 1992). It is also known that, Kernel and Polynomial Regression Estimators provide good estimates for the Population Total when $x \in \mathbb{R}^d$ and $d = 1$, (Otieno and Mwalili, 2000; Montanari and Ranalli, 2003).

However, even though high dimensional Auxiliary Information might be accounted for in the above Estimators, the problem of the sparseness of regressors' values in the design space makes Kernel Methods and Local Polynomials unfeasible as performance deteriorates sharply with increase in the dimension, (Stone, 1982; Bickel and Li, 2007; Montanari and Ranalli, 2003). This problem is known as the **"Curse of Dimensionality"** which is caused by the sparsity of the data in high dimensional spaces, resulting in a decrease in the fastest achievable rates of convergence of the Regression Function Estimators towards their target curve as the dimension of Regressor Vector increases. Local approximators in such a context run into problems. A review on the concept of Curse of Dimensionality is provided in (Friedman, 1991).

Therefore, one has to turn to different Nonparametric Estimators to retain a large degree of flexibility. An attempt to handle Multivariate Auxiliary Information is to use Recursive Covering in Model-Based perspective (Di Ciaccio and Montanari, 2001) and Generalized Additive Modeling in a Model-Assisted Framework (Opsomer et al., 2007). These estimation methods comes at a cost of reduced flexibility with the associated risk of increased bias (Stone, 1982; Friedman, 1991; Bickel and Li, 2007; Rady and Ziedan, 2014).

To address the mentioned challenges, studies have suggested the use of Neural Network in the Estimation of the Mean Function in cases where we have higher dimensional datasets, (Cybenko, 1989; Funahashi, 1989; Barron, 1993; Franke and Diagne, 2006). Neural Network(NN) constitute a class of flexible Nonlinear Models designed to mimic Biological Neural Systems. Biological Neural System consists of several layers, each with a large number of Neural Units (Neurons) that can process the information in a parallel manner. The models with these properties are known as NN Models.

The development in NN, including more complex and flexible NN structures and new Network learning methods was first presented by (Rumelhart et al., 1988) in his seminal work. Since then, NN has become a rapidly growing research, attracting interest in different fields. For instance, it has been applied in Pattern Recognition, Signal Processing, Language Learning and many more in the field of Finance, Econometrics etc.

The reason behind the choice of NN in this work is that NN can be used to approximate the unknown Conditional Mean Function of a variable of interest without suffering from the problem of Model Misspecification, unlike Parametric Models commonly used in empirical Studies. This is because of its Multi Layer Structure in which the Middle Layer is build upon many Simple Nonlinear Functions that play the role of Neurons in Biological Systems. Therefore, by allowing the number of these Simple Functions to increase indefinitely, a Multi Layered NN is then capable of approximating a large class of functions to any desired degree of accuracy as shown in the theoretical work by (Cybenko, 1989; Funahashi, 1989; Barron, 1993).

Although Kernel and Local Polynomial Approximators also have the same property, they usually require a large number of components to achieve similar approximation accuracy (Barron, 1993). NN are thus considered to be a parsimonious approach to Parametric Functional Analysis.

1.2 Statement of the Problem

Consider the estimation of the Population Total say T

$$T = \sum_{i \in s} y_i + \sum_{i \in r} y_i, \tag{1.3}$$

where s are the Sample Units and r the Nonsampled Units. Assume that

$$y_i = m(X_{i,j}) + \varepsilon_i, \tag{1.4}$$

with $X_{i,j} \in \mathbb{R}^d$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ *i.i.d* with mean zero and $x_{ij}, i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, d$ are the Auxiliary Information. For $d = 1$ (One Dimension), Nonparametric Estimates for the Mean Function $m(.)$ based on the Kernel or Local Polynomial Estimates have been shown to provide good Estimates of the Population Total T in equation (1.3), (Stone, 1982; Otieno and Mwalili, 2000). It has been demonstrated that Multivariate Auxiliary Information might be accounted for in the above Estimators, however, the problem known as *Curse of Dimensionality* makes these methods unfeasible as performance deteriorates sharply with increases in Dimension (Stone, 1982; Bickel and Li, 2007; Montanari and Ranalli, 2003). The reason behind this poor performance is because, these methods rely on Local Averaging, therefore, in high dimensions, the Local Neighborhoods are almost empty and the Neighborhoods that are not empty are not Local hence the boundary effect are greatly exaggerated making the Computational and Statistical Efficiency of the estimator difficult. In such cases, one has to either constrain the form of the Mean Function $m(.)$ for instance, use Functions which are Additive with respect to the coordinates of x_j but these comes at a cost of reduced flexibility and an associated risk of increased Bias (Stone, 1982; Friedman, 1991; Bickel and Li, 2007; Rady and Ziedan, 2014). Another alternative is to turn to different Nonparametric Estimators to retain a large degree of flexibility, (Montanari and Ranalli, 2003; Stone, 1982). To address this challenges, this study considers a Feedforward Backpropagation Neural Network learning to Estimate the Functional Relationship between the Survey Variable and the Auxiliary Variables in High Dimensional case.

1.3 Objectives of the Study

1.3.1 General Objective

To estimate Finite Population Total based on a Robust Nonparametric Feedforward Backpropagation Neural Network.

1.3.2 Specific Objectives

1. To develop a Robust Nonparametric Estimator of Finite Population Total based on Feedforward Backpropagation Neural Network.

2. To derive the Asymptotic Properties of the developed Estimator.

3. To compare the performance of the developed Estimator to other existing Nonparametric Finite Population Total Estimators such as Generalized Additive Models, Multivariate Adaptive Regression Spline and Local Polynomial which can handle High Dimensional Data using simulation Procedure.

4. To apply the developed Estimator to the data acquired from the United Nations Development Programme 2020 and compare its performance with other existing Nonparametric Finite Population Total Estimators such as Generalized Additive Models, Multivariate Adaptive Regression Spline and Local Polynomial.

1.4 Justification of the Study

1.4.1 Justification to the Theory of Statistics

The main goal in Survey Sampling is to use the Sample Statistics to make conclusions about the overall Finite Population Quantities. Nonparametric Regression has developed into a increasingly growing field of statistics. This Regression approach is flexible and data-analytical way of Estimating Regression Function without specifying a Parametric Model correctly. Nonparametric Estimates are often more reliable and flexible than Design Based Presumptions or Parametric Regression Models. In Sample Surveys, the Auxiliary Information is used to increase the accuracy of Estimators of Finite Population Quantity at the Estimation Stage.

Within the context of Nonparametric Regression, the Estimator suggested in literature contribute to the Trade-Off of Bias-Variance along the Boundary Points and hence becomes infeasible in high dimensions.The aim of this research is therefore to address this weakness by applying Neural Network method to the Estimation of the Finite Population Total in high dimensional case.

1.4.2 Justification to Users of Statistics and other Stakeholders

Globally, Census plays a crucial role during resource allocation and planning. However, they are carried out only after every ten years. Thus, other methods are required for planning in the intervening years. Population Estimates use the Census as a baseline, for instance adding Births and subtracting Deaths and make allowances for Migration. They can be used for National and Local Planning. Population Estimates are produced annually.

Additionally, National Government use Population Estimates as the basis for capitation-based funding of County Governments, Primary health Care, Education and other sectors of the economy, hence under-estimation can therefore have effects on Local services, and Over-estimation can lead to unfair resource distribution. Therefore having an Robust Estimator Population Total/ Estimates will ensure equitable resource allocation.The outcome of this thesis will play an important role in providing a reliable Estimator of Finite Population Total that will assist the Government in ensuring equitable resource allocation.

Additionally, the study contributes towards development of Mathematical and Statistical knowledge in Survey Sampling. The developed Estimation Procedure is useful to policy makers since National Development is dependent on the Sampling Strategy employed. In addition, Business and Industrial sectors stand to benefit from this study by using the developed Estimation Procedure for prediction and thereby improving the efficiency of their internal operations.

1.5 Organization of the Thesis

The rest of this thesis is organized as follows: In chapter two, a critical review of the work done by other researchers in the Nonparametric Estimation of the Finite Population Parameters is accomplished and also some of the Robust Estimators of the Finite Population Total are reviewed. In chapter three, Neural Network is reviewed extensively and Robust Estimators of the Finite Population Total using the procedure of Neural Network is developed in a Model Based Framework and its properties investigated. In chapter four, a study is carried out to compare the performances of the Estimator developed in chapter three with some other Estimators that exist in the literature. Finally, in chapter five, a Summary of the study is outlined in terms of the Conclusions and Recommendations for Further Research.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

Estimation of a Finite Population Parameters such Population Mean and Total has been an important problem in Survey Sampling. The most common approach to Estimation of Population Parameters in Finite Population Sampling takes a Design Based Approach, hence a number of Design Based Procedures such as Horvitz-Thompson Estimator have been developed for the Estimation of Finite Population Total.

Another widely used approach to this problem involves using any available Auxiliary Information related to the Population of interest that is available. One approach to using this Auxiliary information in Estimation is to assume a working model describing the relationship between the study variable of interest and the auxiliary variables. Estimators are then derived on the basis of this model, (Dorfman, 1992b). Estimators are sought which have good efficiency if the model is true, but maintain desirable properties like design consistency if the model is false.

During such problems, the main challenge that needs to be addressed is the matter of accuracy of the Estimator which may also be referred to as Robustness of the Estimator. Accordingly, each time an Estimator of Finite Population Total is constructed, matters of robustness in the chosen Estimator are of importance. The current problem of this thesis is on the Estimation of Finite Population Total in high Dimensional spaces, a problem which suffers from Robustness due to high biases at the boundaries if an estimator is not correctly chosen. Herein, to ensure Robustness in the Estimation of Finite Population Total in high dimensions (multivariate case), a Nonparametric Feedforward Backpropagation Neural

Network Function is developed and used.

The reason behind the choice of a Nonparametric Feedforward Backpropagation Neural Network is informed by the fact that a Neural Network can be used to approximate the unknown Conditional Mean Function of a variable of interest without suffering from the problem of model misspecification, unlike Parametric Models commonly used in empirical studies. Additionally it does not rely on the local averaging that Kernel Smoother in the Nonparametric context relies on. This is because of its Multi-Layer Structure in which the Middle Layer is build upon many simple Nonlinear functions that play the role of Neurons in Biological Systems.

Therefore, by allowing the number of these simple functions to increase indefinitely, a Multi-Layered NN is then capable of approximating a large class of functions to any desired degree of accuracy as shown in the theoretical work by (Cybenko, 1989; Funahashi, 1989; Barron, 1993). Once a Nonparametric Feedforward Backpropagation Neural Network Function is developed, it is then employed in the Estimation of Finite Population Total.

2.2 Theoretical Review

2.2.1 Nonparametric Estimators of Finite Population Total

In this section, numerous studies in Nonparametric Estimation process have been discussed that formed the basis for the Neural Network Estimator that was developed in the current study.

It is important to note that in Nonparametric Estimation procedures, when the Auxiliary information is available, a Linear Model is selected as the working model, (Dorfman, 1992b). Estimators such as Generalized Regression Estimators, (Cassel et al., 1976; Särndal, 1980; Robinson and Särndal, 1983)), including Ratio Estimators and Linear Regression Estimators, (Cochran, 1977), Best Linear Unbiased Estimators (Brewer, 1963) and Post-Stratification Estimators (Holt and Smith, 1979), are all derived from assumed Linear Models.

In some situations, the Linear Model is not appropriate and the resulting Estimators do not achieve any efficiency gain over purely Design-Based Estimators.

(Wu and Sitter, 2001) propose a class of Estimators for which the working models follow a Nonlinear Parametric Shape. However, efficient use of any of these Estimators requires a priori knowledge of the specific Parametric Structure of the Population. This is especially problematic if that same model is to be used for many variables of interest, a common occurrence in Surveys.

Chambers et al. (1992) considered the same problem of estimating Finite Population Total and proposed a calibrated version of the Nonparametric Estimator. The prediction Bias was estimated based on a specified model and subtracted from the Nonparametric Estimate of the Nonsampled component. The resulting Estimator had a better performance than the Non-calibrated Nonparametric Estimator.

As a way of reducing the Boundary Bias that is experienced when Estimating using Kernel Smoothers, (Herbert et al., 2017) considered the problem of Estimating Finite Population Total and proposes incorporating the Jackknifed procedure into the Nonparametric Regression Estimator (the case of Nadaraya- Watson) to reduce the Bias. The empirical results from the simulations conducted showed that in terms of Biases and Mean Square Errors, this Estimator performed well compared to other existing Estimators in the Estimation of the Finite Population Total.

One problem in the above literature in Estimation of the Finite Population Total is that they have been considered in the Univariate case; that is $X \in \mathbb{R}^d,\ d = 1$. As noted by (Montanari and Ranalli, 2003), extension of these techniques to Multivariate case although it is feasible in theory, it is difficult in practice because of the Curse of Dimensionality. The cause of this *Curse of Dimensionality* is the Trade-Off between the Bias and Variance in Nonparametric Curve Estimation. Bias controls demand to consider data in a small Neighborhood around the target predictor X, where the Curve Estimate is desired, while Variance control requires large Neighborhoods containing many predictor-response pair.

Therefore, when the dimension increases, the predictor location becomes increasingly sparse, with larger average distance between predictor location, moving the Bias-Variance Trade-Off and resulting rate of convergence in unfavorable direction.

(Bickel and Li, 2007) demonstrates that, "naive" Multivariate Local Polynomial Regression can adapt to a Local Smooth Lower Dimensional Structure in the sense that it achieves the optimal convergence rate for Nonparametric Estimation of Regression Function belonging to a Sobolev space when the predictor variables lives on or close to a lower dimensional manifold. However, this will require fitting a model with a large sample size to achieve this. (Stone, 1982) in his study demonstrates that, if the Regression Function $m(x)$ belong to a Sobolev space with smoothness p, there is no Nonparametric Estimator that can achieve a faster convergence rate than $n^{-\frac{p}{2p+d}}, \ where \ d$ is the dimension of the predictor variables.

There has been a surge in research on identifying intrinsic low dimensional structure from seemly high dimensional source, (Stone, 1982; Belkin and Niyogi, 2003; Ham et al., 2004). In this case, it is assumed that the observed high dimensional data are lying on low Dimensional Smooth Manifold. If one can Estimate the Manifold, then it can be expected that procedures which perform as well as if the structure is known even if the low Dimensional Structure obtains only in a Neighborhood of a Point Estimation at that point should be governed by actual rather than ostensible dimension. (Levina and Bickel, 2004), points out that, in predicting Y from X on the basis of training sample, one could automatically adapt to the possibility that the apparently high dimensional X that one observed in fact lived in a much smaller dimensional Manifold and that the Regression Function was Smooth on that Manifold.

Further attempt to handle Multivariate Auxiliary is to use Recursive Covering in Model-based perspective, (Di Ciaccio and Montanari, 2001). Within Model-Assisted Framework, Generalized Additive Models(GAM) have been employed to this end by (Opsomer et al., 2001) who proposed the use of Penalized Splines while (Montanari and Ranalli, 2003) considered Neural Network (NN) in a more general context of Model Calibration to Estimate Population Mean. In their study they found out that NN gain efficiency with respect to Classical Regression Estimators except in case when Sampling in High Dimension. They also observed that, once Weight Decay Penalization is employed, choice of the number of Units in the Hidden Layer is less important and does not imply in any case particularly erratic result.

(Montanari and Ranalli, 2003) also compared the performance of NN, DART, (Friedman, 1991), MARS, (Friedman, 1991) and GAM, (Hastie and Tibshirani, 1990). The theoretical properties of these Estimators were stated and their performance tested through a Simulation Study where they found out that NN also competed well among the other Estimators in the Univariate case. (Wangang et al.2014) explores the use of MARS and NN to capture the Intrinsic Nonlinear and Multidimensional Relationship associated with Pile Driveability. In their study, Performance measures indicated that NN and MARS models for the analyses of Pile Driveability provide similar predictions and can thus be used for predicting Pile Driveability.

From these reviews on Parametric and Nonparametric Estimators of Finite Population Total that have been proposed and used to date, It may be seen that these estimators have shown good performance in situations where the model has been specified correctly for the Parametric ones and in lower dimensions (univariate case) for the case of Nonparametric Kernel Smoothers. Kernel Smoother Estimators generally relies on Local Averaging hence in high dimensions, the boundary effect are greatly exaggerated as the dimension increases since a fraction of data points near the boundary grows rapidly making the computational and statistical efficiency of the estimator difficult.

Therefore, it comes out clearly that where the Model is Misspecified, the Parametric Estimators of Finite Population Total discussed in the Literature above will give Estimators that are not Robust and are of low precision, while the Estimators based on Kernel Smoothers their efficiency will become poorer as the dimension of the Regressors increases. Therefore, this calls for the need of an Estimator of Finite Population Total that can handle issues of model Misspecification and maintain high efficiency in High Dimensional datasets.

Thus, in this thesis, an estimator of Finite Population Total based on a Robust Nonparametric Feedforward Backpropagation Neural Network is developed.

2.2.2 Asymptotic Properties of Estimators Based on Nonparametric Regression

The key properties that a statistician would be interested to check given an Estimator, are the, Normality, Consistency, the Variance and the Bias of that Estimator. These can enable one to measure the amount of precision and accuracy that an Estimator has. In fact at an arbitrary fixed point, a basic measure of accuracy that takes into account both the Bias and Variance is the Mean Square

Error (MSE) (Tsybakov and Tsybakov, 2009).

Other texts that have such literature include (Härdle and Linton, 1994; Härdle et al., 2004; Takezawa, 2005). This is one of the criteria of error measurement that can be used in such statistical researches. In Nonparametric Regression Estimation, one may be interested in the cumulative amount of Bias and the Variance over the entire regression line. This global measure called MISE is obtained by finding the integral value of the Variance and the Square of the Bias of the Estimator (Zucchini et al., 2003; Soh et al., 2013).

An asymptotic approximation of Univariate Kernel Estimator using Taylor's series expansion will yield the Asymptotic Mean Integrated Square Error (AMISE), (Manzoor et al., 2013). Given the Asymptotic Properties one can discuss the speed of convergence of the Estimators and determine the cost to pay in a given option. It is from this vast literature that this study uses these measures in the analysis stage to compare the proposed estimator against the standard ones reported in the simulation study.

From these reviews it can be seen that determining Asymptotic Properties of an Estimator play a key role in the development of that Estimator. Therefore, as the one of the stages of coming up with estimator of Finite Population Total in high Dimensions cases, the study has to derive the Asymptotic Properties as an indicator and measure of performance.

2.3 Empirical Review

2.3.1 Performance of Estimators of Finite Population Total

The performance of an Estimator or a model is related to how close are the prediction values to the observed values. In assessing the performance of Estimators, measures that allow for comparison of the Estimators are normally subjected to simulated data. There are different consistency criteria used in order to compare and assess the performance of different Estimators. This includes Mean Square Error (MSE), Mean Absolute Error (MAE), Bias, Relative Bias (RB), Mean Absolute Percentage Error (MAPE) and the list continues.

(Dorfman, 1992b) used the Root Average Relative Biases to make comparison between the Design-Based Hovitz-Thompson Estimator with the Model-Based Nonparametric Regression Estimator derived using (Nadaraya, 1964) Smoother. More literature on the techniques used to measure the performance of the Estimators in different studies can be found in (Bickel and Li, 2007; Breidt and Opsomer, 2000; Breidt et al., 2005; Otieno and Mwalili, 2000; Otieno et al., 2007). From the above literature, it is imperative to note that, performance of an Estimator is an important step in checking whether the developed Estimator can be reliable is the Estimation of particular Population Parameters. The measures used in these literature are universal and employed in most of the Statistical Analysis to compare the performance of different Estimators. These measures are also employed in this study to compared the developed Estimator with the other identified Estimators.

2.4 Research Gaps Resulting from Critiques of the Existing Literature Reviewed

From the literature reviewed in Sections 2.2.1, 2.2.2 and 2.3.1 where the critiques have been given, it was clear that Parametric Estimators and Nonparametric Kernel Smoothers are popular in the Estimation Finite Population Parameters such as Population Total. They have gained usage in many practical cases and relied upon by many researchers. Various Nonparametric Estimation methods of Finite Population Total reviewed employed either Kernel Smoothers, Local Polynomials and the Splines in the Estimation of the Regression Function which is later used to predict the Nonsampled Units in Population.

However, conventional Parametric Estimators suffer when the model under consideration is Misspecified and can not capture Non-Linearity in the data which often exit when dealing with real life situations, while most Kernel Smoothers have boundary problems and their efficiency rely greatly on the bandwidth that is selected. In addition, for the Kernel Smoother, Estimators have been shown to only provide good Estimates of the of the Regression Function in lower dimensions but as the dimension increases, their performance deteriorates.

Therefore, there is still no Robust Estimation Framework in Nonparametric Estimation techniques that can be used to efficiently Estimate the Regression Function and hence the Finite Population Total in cases of high dimensional dataset.

This study uses Feedforward Backpropagation Neural Network approach to the Nonparametric Estimation of Finite Population Total.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

In real-life, Surveys are conducted at Local, National and International levels to gather information and aid Public and Private sectors in effective policy making. The reason behind any Survey is usually to obtain summary statistics for a Finite Population or for specific subgroups at the same time reducing the time and cost of collecting data. In this instance, the problems involving Estimation of Population Parameters such as Population Total, Means, and Proportions will occur.

To find an efficient Estimator of one these Parameters, one has to first evaluate its properties of the Estimators available and then choose the best one. The most desirable properties are typically Unbiasedness, Minimum Variance, Consistency and least Mean Square Error. It is worth noting that using a Survey approach can help build an Estimator with these desirable properties. Careful application of such methods often achieves better performance. Most researches have developed Estimators that works well when we have small dimensional datasets especially in the Univariate case.

The truth is, in real life situations , we experience problems that rely on Multivariate datasets or what we refer in here as the high dimensional case and therefore the existing Estimators face a challenge in the estimation of population parameters such as Population Total with high precision. Therefore, this Estimation of Finite Population Total calls for a Robust Estimator which can handle high dimensional datasets and still maintains high precision.

In addition, it has been observed from literature that, Nonparametric Regression techniques based on Kernel and Local Smoothers suffer greatly from the problem of the Curse of Dimensionality hence becoming unfeasible as performance deteriorates sharply with the increase in dimension. Also, if one resorts to using Parametric Estimators in the Estimation, they also suffer from Model Misspecification and also Multicollinearity of the Regressors which affects their performance and inability to capture Non-linearity in the data.

The alternative is to use Neural Network (NN) in the Estimation of the Regression Function. Neural Network is able to handle high dimensional data because of its Multi-layered structure in which the Middle Layer is made up of Neurons. Therefore, allowing number of Neurons to increase indefinitely enables NN model to approximate the unknown Conditional Mean Function of a variable of interest without suffering from the problem of Model Misspecification and also gives it the capability of approximating a large class of functions to any desired degree of accuracy.

To enable development of a Robust Feedforward Backpropagation Neural Network Estimator of Finite Population Total, a brief review of the general theory of Neural Network is now done in the Section 3.4 that follows. Here we review the Activation Functions with an intention of highlighting reasons towards choice of the Logistic Activation Function that is eventually used in this development. The reason behind the choice of the Backpropagation as a training technique is also provided here. Finally, some Notations used are provided and briefly explained.

3.2 Notations Used

In this section, the fundamental concepts and Notations useful for the sequel are provided and defined.

Suppose there is a Finite Population of N with distinct and identifiable Units $U = \{1, 2. \ldots, N\}$. We define the Population Total as

$$T = \sum_{i=1}^{N} Y_i \tag{3.1}$$

$$T = \sum_{i \in s} y_i + \sum_{i \in r} y_i \tag{3.2}$$

The intention is to Estimate the Nonsampled elements in the Population r. Let there exist an Auxiliary variable $X \in \mathbb{R}^d$, closely associated with Y, which is known for the entire Population ($i.e\ X_1, X_2 \ldots, X_N$) are known $\forall\ Y_i$. Let X and Y be define by a Superpopulation Model such that

$$y_i = m(x_{i,j}) + \varepsilon_i \tag{3.3}$$

with $x_{i,j} \in \mathbb{R}^d$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ $i.i.d$ with mean zero and $x_{ij}, i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, d$ are the Auxiliary Information. The the Population Total T can be given as

$$T = \sum_{i \in s} y_i + \sum_{i \in r} m(x_{i,j}) \tag{3.4}$$

Where $m(x_{i,j})$ is the Mean Function which is Estimated by Neural Network Function given as

$$f_H(x, \theta) = v_0 + \sum_{h=1}^{H} v_h \psi\left(w_{0h} + x^T w_h\right),\ x \in \mathbb{R}^d \tag{3.5}$$

where θ represents the a vector of Neural Network Weights, H denotes the number of nodes/ neurons, v_h denotes the Network Weights from the hidden layer to the output layer, w_h denotes the Network Weights from the input layer to the Hidden layer and ψ represents the Activation Function.

3.3 Nonparametric Estimators of Finite Population Totals

3.3.1 The Nadaraya-Watson Estimator

In Estimating the Population Total in equation (1.3), a datum point remote from x carries very minimal information about $m(x)$. The Estimator is therefore Estimated using the Function of the Sample values of $y_i's$ and the Nonsample component is predicted, based on the Nonparametric model in equation (1.4). The non-sample values of $y_j's$ are estimated using the local running average such that

$$\hat{m}(x_j) = \sum_{i=1}^{n} w_i(x_j) y_i \tag{3.6}$$

where $w_i(x_j) = \dfrac{k\left(\frac{x_i - x_j}{h}\right)}{\sum\limits_{i=1}^{n} k\left(\dfrac{x_i - x_j}{h}\right)}$ is the Nadaraya- Watson Smoother (Nadaraya, 1964). The Function $k(u)$ is a Symmetric Density Function, it can be the Gaussian, Rectangular, Triangular, biweight or Epanechnikov Kernel for given Scaling Smoothing Parameter h. The Kernel Function is defined by the relation given by

$$k(u) = \frac{1}{nh} k\left(\frac{x_i - x_j}{h}\right)$$

The Kernel Function $k(u)$ is under the user control. Therefore, it is necessary for practical purposes to consider results that hold for particular Kernel being used. The assumption considered at this stage is that the Kernel is Symmetrical Function satisfying the following properties, (Silverman, 1986);

i) $k(u) \geq 0$

ii) $\int k(u)du = 1$

iii) $\int uk(u)du = 0$

iv) $\int u^2 k(u)du = K_2 \neq 0$

v) $k(u) = k(-u)$ for all u

vi) $\int_{-\infty}^{\infty} (k(u))^2 du < \infty$

It has been noted that none of the smoothing procedures is uniformly best. However, the Kernel Smoothers have optimal properties, (Gasser and Engel, 1990). A Nonparametric Regression Estimator for the Finite Population Total, (Dorfman, 1992a,b) takes the form

$$\hat{T}_{NW} = \sum_{i \in s} y_i + \sum_{j \in r} \hat{m}(x_j) \tag{3.7}$$

It can be shown that the Conditional Mean and Variance of $\hat{T}_{NW} - T$ can be give as follows

$$E(\hat{T}_{NW} - T) = \frac{(N-n)h^2 K^2}{2} \int \beta(x) d_s^{-1} d_r(x) dx + O(nh^3 + n^{\frac{1}{2}} h^{\frac{1}{2}})$$

$$
\begin{aligned}
V(\hat{T}_{NW}-T) = \frac{(N-n)^2}{n} & \int \sigma^2(x) d_x^{-1}(d_r(x))^2 dx + \\
& \frac{(N-n)^2}{nh} \int K^2(x) du \int \sigma^2(x) d_s^{-1}(x) d_r(x) dx + \\
& \frac{(N-n)^2}{n} h^2 K_2 \int C^*(x) dx + \\
& (N-n) \int \sigma^2(x) d_r dx + O(nh^2 + n^{\frac{1}{2}} h^{\frac{1}{2}})
\end{aligned}
$$

where $C^*(x)$ is a complicated function of the derivatives of $d_s(x)$ and $d_r(x)$. and the Mean Square Error if given by

$$
\begin{aligned}
MSE(\hat{T}_{NW}) = \big[& h^2(N-n)^2 K_2 \\
& \int \beta(x) d_s^{-1} d_r dx + o(nh^2 + h^{-1}\big)^2 \\
+ \frac{(N-n)}{n} & \\
& \int \sigma^2(x) d_s^{-1}(x) \left[d_r(x)\right]^2 dx + (N-n) \\
& \int \sigma^2(x) d_r(x) d(x) + o(x)
\end{aligned}
$$

For proof of these Asymptotic Properties, see (Dorfman, 1992a; Githinji.S, 2010).

In higher dimensions, $X_i \in \mathbb{R}^d$, the Kernels can easily be used by just replacing $x_i - x$ in the Kernel argument by $\parallel x_i - x \parallel_2$ so that the Multivariate Kernel Regression Estimator is

$$
\hat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{\parallel x_i - x \parallel_2}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{\parallel x_i - x \parallel_2}{h}\right)}
$$

and the corresponding Estimator for Finite Population Total will be

$$
\hat{T}_N = \sum_{i \in s} y_i + \sum_{j \in r} \hat{m}(x_j) \tag{3.8}
$$

(Godambe, 1955; Dorfman, 1992b), considered a problem of Estimating Finite Population Total using Nonparametric Regression. In their work, they considered this Nadaraya-Watson Estimator to Estimate the Mean Function to predict the Nonsampled values of the study variable and consequently to Estimate the Finite Population Total. It was demonstrated that, as long as some standard conditions were met, the Relative Bias of the above Estimator goes to zero.

The ratio of the Bias to the Standard Error of the Estimator was also shown to be Asymptotically zero suggesting that using a wide bandwidth might yield a better Estimate and it was found to be more efficient when compared to rival Design Based Estimators. The greatest efficiency was achieved when a Linear Model was used and Variance assumed to be proportional to the square of the Auxiliary variable. Just like in the Univariate case where this Estimators suffers from the Boundary effects, (Franke and Diagne, 2006), noted that, this Kernel Smoother Estimator in high dimensions, the boundary effect are greatly exaggerated as the dimension increases since a fraction of data points near the boundary grows rapidly making the computational and statistical efficiency of the Estimator difficult.

3.3.2 The Gasser-Muller Estimator

The denominator in the Nadaraya-Watson Estimator (Nadaraya, 1964) is convenient when taking derivatives of the Estimator and when deriving its Asymptotic Properties. The sorting X-variable and the Estimator were proposed, (Gasser and Müller, 1979). The Estimator $\hat{m}(x)$ is given by

$$\hat{m}(x) = \sum_{j=1}^{n} \int_{s_{j-1}}^{s_j} k\left(u - x\right) du S_j \tag{3.9}$$

where $S_j = \frac{1}{2}\left(x_j + x_{j+1}\right),\ x_0 = -\infty \ and \ x_{n+1} = +\infty$
Therefore, the corresponding Nonparametric Estimator for the Finite Population Total in equation 1.3 is

$$\hat{T}_G = \sum_{i \in s} y_i + \sum_{j \in r} \hat{m}(x_j) \tag{3.10}$$

This Kernel Smoother Estimator in high dimensions just like the Nadaraya Watson Estimator considered in (Dorfman, 1992a), this Estimator relies on Local Averaging hence the boundary effect are greatly exaggerated as the dimension increases since a fraction of data points near the boundary grows rapidly making the computational and statistical efficiency of the Estimator difficult.

The Priestly-Chao Estimator

(Priestley and Chao, 1972) proposed an Estimator for the unknown Mean Function both when the observations are assumed to be taken at equally spaced in-

tervals and in a case where this restriction is removed.
The Priestly-Chao has the relation $w_i(x_j) = \left(\frac{x_i - x_{i-1}}{h}\right) k \left(\frac{x_i - x_j}{h}\right)$ such that the estimator Priestley and Chao (1972) is given by

$$\hat{m}(x_j) = \frac{1}{nh} \sum_{i \in s} w_i(x_j) y_i \tag{3.11}$$

This Smoother has the disadvantage where one needs to Estimate or extrapolate values of the independent variable. This is another task that will lead to increase of the errors in the Estimation if not correctly determined. Unlike the usual weighting, the sum of weights does not equal to one but is only an approximation.

The Priestley-Chao and Gasser-Muller, (Priestley and Chao, 1972), assume that the Estimator of data is ordered according to the Auxiliary variable data such that $x_{i-1} < x_i$ and their weights are only applicable to cases in which the auxiliary variable is restricted to some interval $[0, 1]$. The Population Total Estimator is given by

$$\hat{T}_C = \sum_{i \in s} y_i + \sum_{j \in r} \hat{m}_C(x_j) \tag{3.12}$$

Since the estimator is based on the Kernel Smoothers, its performance also deteriorates as the dimension of the explanatory variable increases.

3.3.3 The Spline Estimator

Spline Functions are more attractive due to their flexibility and less vulnerability to the Bias resulting from Model Misspecification. In Spline Estimation the technique of Residual Sum of Squares is used (Härdle and Stoker, 1989)

$$\hat{m}(x) = \sum_{i=1}^{n} (y_i - g(x_i))^2$$

Here $g(x)$ is a curve unrestricted in the Functional form. The distance can be reduced by any $g(x)$ that interpolates the data. The technique has disadvantage such that the curve is not unique and too wiggly for a structure-oriented interpolation. The technique has good results because it will produce a good fit to the data and the curve does not have too much rapid local variation. The Spline

Estimator for the Population Total see (Zheng and Little, 2003), is given by

$$\hat{T}_S = \sum_{i \in s} y_i + \sum_{j \in r} \hat{m}(x_j) \tag{3.13}$$

In this context, (Zheng and Little, 2003) considered a model-based alternative to the Horvitz Thompson estimator that employs Penalized Spline Regression.

3.3.4 Multivariate Adaptive Regression Spline (MARS)

$MARS$ was first proposed by (Friedman, 1991) as a flexible procedure to organize relationships between a set of input variables and the target dependent that are nearly additive or involve interactions with fewer variables. It is a Nonparametric Statistical Method based on a divide and conquer strategy in which the training data sets are partitioned into separate piece-wise Linear Segments (Splines) of differing gradients (slope). MARS makes no assumptions about the underlying Functional Relationships between dependent and independent variables.

In general, the Splines are connected Smoothly together, and these Piece-Wise Curves (polynomials), also known as Basis Functions (BFs), result in a flexible model that can handle both Linear and Nonlinear behavior. The connection/interface points between the pieces are called Knots. Marking the end of one region of data and the beginning of another, the candidate Knots are placed at random positions within the range of each input variable.

MARS generates BFs by step-wise searching overall possible Univariate candidate Knots and across interactions among all variables. An Adaptive Regression Algorithm is adopted for automatically selecting the Knot Locations. The $MARS$ algorithm involves a Forward Phase and a Backward Phase. The Forward Phase places candidate Knots at random positions within the range of each predictor variable to define a pair of BFs. At each step, the model adapts the Knot and its corresponding pair of BFs to give the maximum reduction in sum-of-squares Residual Error. This process of adding BFs continues until the maximum number is reached, which usually results in a very complicated and over-fitted model. The backward phase involves deleting the redundant BFs that made the least contributions.

Let y be the target dependent responses and $X = (X_1, X_2, \ldots, X_d)$ be a matrix of d Input Variables. Then it is assumed the data are generated based on an unknown model. For a continuous response, this would be:

$$y = f(X_1, X_2, \ldots, X_d) + \varepsilon = f(X) + \varepsilon \tag{3.14}$$

in which ε is the Fitting Error. f is the built $MARS$ model, comprising of BFs which are Splines Piece-Wise Polynomial Functions. For simplicity, only the Piece-Wise Linear Function is expressed and considered in this study. Piece-Wise Linear Functions follow the form $max(0, x - t)$ with a Knot defined at value t. Expression $max(.)$ means that only the positive part of (.) is used otherwise it is assigned a zero value. Formally,

$$max(0, x - t) = \begin{cases} x - t, \ if \ x \geq t \\ 0, \ otherwise \end{cases} \tag{3.15}$$

The MARS model f(X), which is a Linear Combination of BFs and their interactions, is expressed as

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m \lambda_m(X) \tag{3.16}$$

where each λ_m is a BF. It can be a Spline Function, or interaction BFs produced by multiplying an existing term with a truncated linear function involving a new/different variable (higher orders can be used only when the data warrants it; for simplicity, at most second order is adopted). The term β is constant coefficients, estimated using the least-squares method.

The $MARS$ modeling is a data-driven process. To construct the model in equation 3.16, first the forward phase is performed on the training data starting initially with only the intercept β_0. At each subsequent step, the basis pair that produces the maximum reduction in the training error is added. Considering a current model with M basis functions, the next pair to be added to the model is in the form of

$$\hat{\beta}_{M+1}\lambda_1(X)max(0, X_j - t) + \hat{\beta}_{M+2}\lambda_1(X)max(0, t - X_j) \tag{3.17}$$

with each β being estimated by the least-squares method. This process of adding BFs continues until the model reaches some predetermined maximum number, generally leading to a purposely over fitted model.

The Backward Phase improves the model by removing the less significant terms until it finds the best Sub-Model. Model subsets are compared using the less computationally expensive method of Generalized Cross-Validation (GCV). The GCV is the Mean-Squared Residual Error divided by a penalty that is dependent on Model complexity. According to (Hastie and Tibshirani, 1990), for the training data with N observations, then GCV is calculated as

$$GCV = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[Y_i - f(X_i)\right]^2}{\left[1 - \frac{M+p\times(M-1)}{N}\right]^2} \tag{3.18}$$

in which M is the number of BFs, p is a penalty for each Basis Function included in the developed sub-model, N is the number of datasets, and $f(X_i)$ denotes the $MARS$ predicted values. Thus the numerator is the Mean Square Error of the evaluated model in the training data, penalized by the denominator which accounts for the increasing variance in the case of increasing model complexity. Note that $(M-1)/2$ is the number of Hinge Function Knots. The GCV penalizes not only the number of BFs but also the number of Knots.

After Estimating the function in 3.16 $\hat{f}(X) = \hat{m}(x)_{MARS}$ then the Finite Population Total will be estimated using the function

$$\hat{T}_{MARS} = \sum_{i\in s} y_i + \sum_{i\in r} \hat{m}(x_i)_{MARS} \tag{3.19}$$

For a detailed review of the $MARS$ and its application in Estimation of Surveys Population Parameters see Friedman (1991); Montanari and Ranalli (2003, 2005)

3.3.5 Generalized Additive Models

One of the most popular and useful tools in data analysis is the Linear Regression Model. It is a statistical technique used for Modeling and Analysis of numerical data consisting of values of a dependent variable and of one or more independent variables.

Let Y be a dependent (response) variable, and $X_1, \ldots, X_d$ be d independent (predictor or regressors) variables. To describe the dependence of the mean of Y as a function of $X_1, \ldots, X_d$, it is assumed that the mean of Y is a Linear Function of $X_1, \ldots, X_d$, such that;

$$\begin{aligned}\mu_{y|x} &= E(Y|X_1, \ldots, X_d) = f(X_1, \ldots, X_d) \\ &= \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d \\ &= \beta_0 + \sum_{j=1}^{d} \beta_j X_j\end{aligned} \tag{3.20}$$

Given a Sample of values for Y and X, the Estimates of $\beta_0, \beta_1, \ldots, \beta_d$ are often obtained by the Least Squares Method. It is achieved by fitting a Linear Model which minimizes $\sum_{j=1}^{d} \left(y_j - \hat{f}(x_j)\right)^2$ where $\hat{f}(x_j) = \hat{\beta}_j X_j$

The Generalized Additive Model replaces $\sum_{j=1}^{d} \beta_j X_j$ with $\sum_{j=1}^{d} f_j(X_j)$ where f_j is an unspecified Nonparametric Function. It can be in a Nonlinear form

$$\begin{aligned}E(Y|X_1, \ldots, X_d) &= f(X_1, \ldots, X_d) \\ &= f_0 + f_1(X_1) + \ldots + f_d(X_d) \\ &= f_0 + \sum_{j=1}^{d} f_j(X_j) = m(x)\end{aligned} \tag{3.21}$$

This function$\hat{f}_j(X_j)$ is Estimated in a flexible manner using cubic spline smoother. Thus the Finite Population Total based on the GAM will be estimated as

$$\hat{T}_{GAM} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{m}(x_i)_{GAM} \tag{3.22}$$

3.4 Neural Network

3.4.1 Background

Neural Network is a Network of interconnected Non-linear Processing Units for distributed, parallel processing of input values to obtain a set of output values. It is based on the attempt to model the way a Biological Brain processes data and thus different from other standard Regression Models. Individual Survey attributes can be considered as dependent(output) variables related by a Neural Network to independent variable (input) variables.

The connection weights characterizing the strength" of the interconnection in an

NN corresponds to the Parameters in Multivariate Models. They can be obtained (Estimated) from a Sample of the Population by training. The trained NN can then be used to Estimate the Nonsampled Units of the Survey data which interns can be aggregated to Estimate the Population Total. Numerous studies have indicates that NN Models represents a promising modeling technique especially for data sets having Nonlinear Relationship, (Asnaashari et al., 2013).

This study adopts a FeedForward Neural Network ($FFNN$) in which no feedback is allowed. The three layers are totally connected and there is no link between units belonging to the same layer. Hidden Units are the processing units which do not receive Exogenous input or deliver final output, but receive inputs from and pass on Output to other units in the Network. The signal in $FFNN$ only travel from inputs to outputs. The Activation in the Hidden Layer is normally preferable to be Non-Linear in such way that each output is a Non-Linear combination of the Linear combination of the inputs and it should be differentiable to facilitate the training process. This activation restricts the amplitude to a preferred closed range between $[0, 1]$ or $[-1, 1]$ since it tend to increase the stability of the Network while learning and also it is useful to maintain the normalization of the input data.

Another vital component in creating a NN Model is the number of Hidden Nodes, which defines the complexity of the model developed. If the number of Neurons of in a hidden layer are too few, NN will not be able to model the data accurately. On the other hand if the number of Neurons in a Hidden layer are too large, it can sometimes be beneficial, but may lead to over-fitting, (Despagne and Massart, 1998).

Determining the Network Architecture is a fundamental task in a NN Model development, (Maier and Dandy, 2000; Farrell et al., 2021). It requires the selection of the optimum number of layers and the number of Nodes in each of the layers. There is no integrated theory for the determination of an optimal NN architecture, but it is generally achieved by fixing the number of layers and choosing the number of Nodes in each layer, (Farrell et al., 2021). For a traditional series estimator (such as Splines) the two choices for the practitioner are the basis (the Spline shape and degree) and the number of terms (Knots), commonly referred to as the smoothing and tuning parameters, respectively. In Kernel Regression, these would respectively be the shape of the Kernel (and degree of Local Polynomial) and the bandwidth(s).

For Neural Network, the same phenomena is present, the Architecture as a whole (the graph structure and activation function) are the Smoothing Parameters while the width and depth play the role of tuning Parameters. The Architecture plays a crucial role in that it determines the approximation power of the Network, and it is worth noting that because of the relative complexity of Neural Network, such approximation and comparisons across Architectures are not simple. At a glance, it may not be clear what function class a given Network Architecture (width,depth, graph structure, and Activation Function) can approximate.

Just as for Classical Nonparametrics, for a fixed Architecture it is the tuning Parameters that determine the rate of convergence. Therefore, choosing the number of Hidden layers is another puzzle in a NN structure. The number of Hidden Nodes allow a Neural Network to capture Nonlinear patterns and detect complex relationship in the data. However, Network with too many Hidden Nodes may cause over fitting problems, leading to poor forecasting ability.

Previous research shows that, one Hidden layer is sufficient to approximate any continuous function, provided that sufficient connection weights are given, (Cybenko, 1989; Funahashi, 1989; Shahin et al., 2002). Because of this, the study will adopt a one layer of Hidden Nodes to help reduce computation time and danger of over-fitting.

Figure 3.1 is an Input-Output map, which has d input nodes, one layer of H Hidden Nodes and an Activation Function $\psi(x)$. The input at Hidden layer Nodes are connected by Weights W_{hj} for $h \in (1, 2, 3, \dots, H)$ and $j \in (1, 2, \dots, d)$ where W_{h0} is the Bias for the i^{th} Hidden Node. The Hidden and Output layers are connected by Weights v_h for $h \in (0, 1, 2, \dots, H)$. Considering an Input Vector $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ the Input $V_h(x)$ to the h^{th} Hidden Node is the value

$$V_h(x) = W_{h0} + \sum_{i=1,j=1}^{n,d} W_{hj} x_j \tag{3.23}$$

where the Weights $W_{h,j}\ for\ j = 0, 1, \dots, d$ in this equations corresponds to the Kernel and Local Polynomial Weights in the Kernel and Local Polynomial Smoothers. The number of Hidden Nodes (Neurons) corresponds to the Bandwidth in the Kernel and Local Polynomial Smoothers.

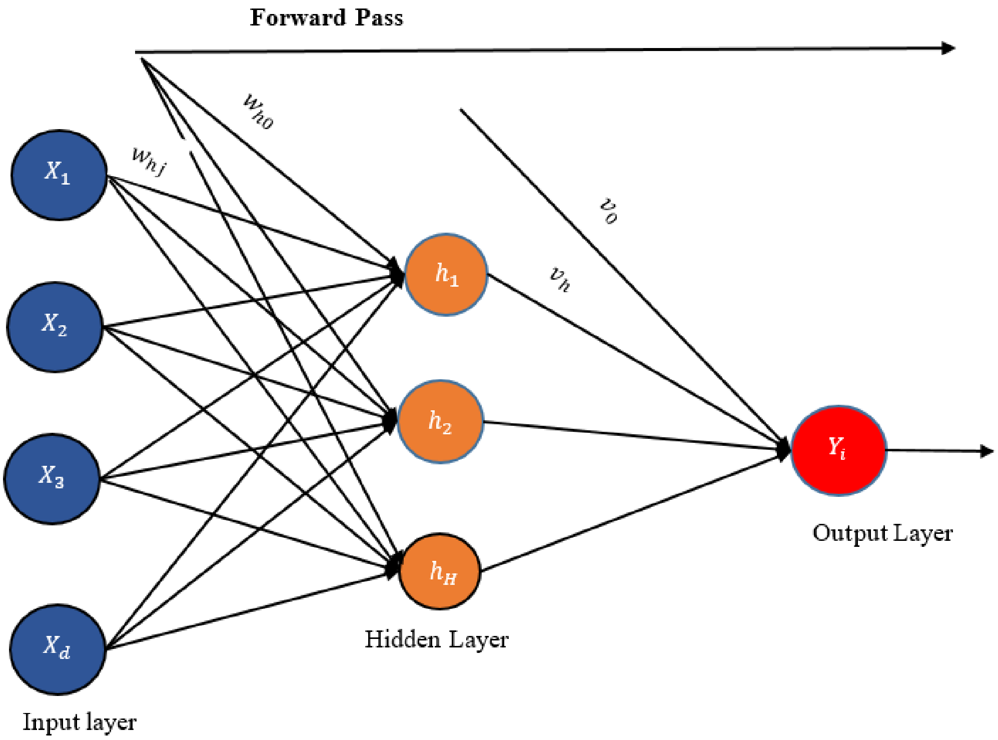

Figure 3.1: Feed Forward Neural Network Structure

The output $\phi_h(x)$ of the h^{th} Hidden Node is the value

$$\phi_h(x) = \varphi(V_h(x)) \tag{3.24}$$

The net Input to the Output Node is the value

$$z(x) = v_0 + \sum_{j=0}^{H} v_h \phi_h(x) \tag{3.25}$$

where $v_h \; for \; j = 0, 1, \ldots, H$ denotes the connection weights from the input layer to output layer. Finally the Output $Z(x)$ of the net is the value

$$Z(x) = \psi(z(x)) \tag{3.26}$$

The connection/ Weights are adjusted through training. There exists two training paradigms: Non Supervised and Supervised training. We discuss and later apply supervised learning. The Supervised training of a Neural Network requires the following: A Sample of d Input Vectors , $X = (x_1, x_2, \ldots, x_d)$ of size n, an associated Output $Y = (y_1, y_2, \ldots, y_n)$ and the selection of an initial Weight set.

We need a repetitive algorithm to update the current Weights to optimize the Input Output map.

3.4.2 Training of the Neural Network

Because of the Nonlinear nature of the Neural Network model for which no direct Estimate exists, an iterative Estimation of the Weights need to be adopted. Training an NN is an optimization problem, where one seeks the minimum of an error surface in a Multi-dimensional space defined by the adjustable Parameters. Such surfaces are characterized by the presence of several Local Minima, Saddle Points. It must be accepted that the NN will probably not find the absolute Minimum of the Error Surface, but a Local Minimum relatively close to the absolute Minimum and acceptable for the problem considered.

There are two methods used to train a Neural Network, the Maximum Likelihood Estimator (MLE) Method and the Sum of Squared Errors (SSE) Method.
The SSE is used to train FeedForward Network. In this methods the Weights are adjusted in such a way that the SSE between the targets Y and the Output Z is Minimized. This SSE is defined as:

$$\begin{aligned} S^2(x_i,\theta) &= \sum_{i=1}^{n} (Y_i - Z(x_i;\theta))^2 \\ &= \sum_{i=1}^{n} (Y_i - Z\left(x_i; W, \alpha\right))^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \psi \left(v_0 + \sum_{j=1}^{H} (v_n \psi(W_{h0} + \sum_{i=j}^{H} W_{hjx_j})) \right) \right)^2 \end{aligned} \tag{3.27}$$

The training process of the Network involves updating the Weights until the function in equation (3.27) is Minimized. There are various methods of Minimizing this function; Backpropagation, Quasi-Newton Method, and Simulated-Annealing Method. In this study, we consider Backpropagation for training our Network but one can consider other methods.The advantages of using a Backpropagation algorithm in this study is that, it does not have any parameters to tune except for the number of inputs. and it is highly adaptable and efficient and does not require any prior knowledge about the network. It is a standard process that usually works well.

Backpropagation (BP)

The BP procedure relies on the differences between the output/ Estimated variable values and the target variable values (the observed values) from the training set as a basis for adjusting weight values in each iteration. The training process can be monitored by watching the computed MSE of all differences after each training cycle. Optimal Weight values are obtained when all the differences between Estimated values approaches zero. In actual applications, the iteration has to be terminated when no further progress in diminishing the differences between Estimated and target values can be observed.

The set of Weights or Parameter values found when the training is terminated may not represent the best or optimal set but a set representing a so called Local Minimum in the MSE surface. A larger number of Weights in the models indicates a danger of over-fitting. Over-fitting is adjusting too much to the training sample with a risk of losing the model ability to generalize and make useful prediction of the Nonsampled Units.

Taking a Unipolar Activation Function $\psi(x)$, the Weights are adjusted as follows

$$\begin{aligned} W^{r+1} &= W^{r} + \Delta W \\ v^{r+1} &= v^{r} + \Delta \alpha \end{aligned}$$

Taking individual Weights, we have the r^{th} iteration Weights as

$$\begin{aligned} v_h^{r+1} &= v_h^r - \lambda_1 \left\{ \frac{\partial S^2(x_i;\theta^{(r)})}{\partial v_h} \right\} \\ &= v_h^r + \lambda_1 \left\{ Y_i - Z(x_i;\theta^{(r)}) \right\} Z(x_i;\theta^{(r)}) \left\{ 1 - Z(x_i;\theta^{(r)}) \right\} \phi_h(x_i) \end{aligned} \tag{3.28}$$

for $i = 1, 2, \dots, n$ and $h = 1, 2, \dots, H$
Similarly;

$$\begin{aligned} W_{hj}^{r+1} &= W_{hj}^r - \lambda_2 \left\{ \frac{\partial S^2(x_i;\theta^{(r)})}{\partial W_{hj}} \right\} \\ &= W_{hj}^r + \lambda_2 \left\{ Y_i - Z(x_i;\theta^{(r)}) \right\} Z(x_i;\theta^{(r)}) \left\{ 1 - Z(x_i;\theta^{(r)}) v_h^r \right\} \\ &\quad - \left\{ \phi_h(x_i) \left(1 - \phi_h(x_i) \right) \right\} x_j \end{aligned} \tag{3.29}$$

for $i = 1, 2, \dots, n$, $h = 1, 2, \dots, H$ and $j = 1, 2, \dots, d$
λ_1 and λ_2 represents the stop gain. The Weights are adjusted until the stopping criterion are met. Under this method, each Weight is adjusted n, the sample size

times each iteration; which implies that for I iterations each Weight is adjusted I_n times.

3.4.3 Choices of Activation Function

For Model Specifications, the building blocks of a NN Model are the Activation Functions ψ. Different choices of the activation functions results in different Networks Models. Here, we look at some Activation Functions commonly used in empirical studies.

The Hidden Units plays the role of Neurons in Biological Systems. Thus, the Activation Function in each Hidden Unit determines whether a Neuron should be turned on or off. Such on/off response can easily be represented using an indicator(threshold) function, also known as Heaviside Function in the NN literature.

$$\psi(u) = \begin{cases} \psi(u) \longrightarrow 0, \ as \ u \longrightarrow -\infty \\ \psi(u) \longrightarrow 1, \ as \ u \longrightarrow +\infty \\ \psi(u) + \psi(-u) = 1 \end{cases} \tag{3.30}$$

In NN literature, it is common to choose a Sigmoid (S-shaped) and Squashing (bounded) Functions because of its characteristics of allowing Nonlinearity and also being differentiable. Another advantage is that the derivatives of the Sigmoid Function can be expressed in terms of the individual Functions itself which is useful when training the Network, (Zilouchian, 2001). Depending on the required output, one could choose between widely used Sigmoid Functions, the Logistic Sigmoid and the Bipolar Sigmoid. The Logistic Function is preferable when the objective is to approximate functions that maps into Probability Space.

In particular, if the input signals are "squashed" between zero and one, the Activation Function is understood as a Smooth counterpart of the Indicator Function. A leading example of the Logistic Function is described as

$$\psi(u) = \frac{1}{1 + \exp(-u)}, \ -\infty < u < \infty \tag{3.31}$$

which approaches one(zero) when its arguments goes to infinity(negative infinity). Thus, the Logistic Activation Function generates a partially on/off signals based on the received input signals.

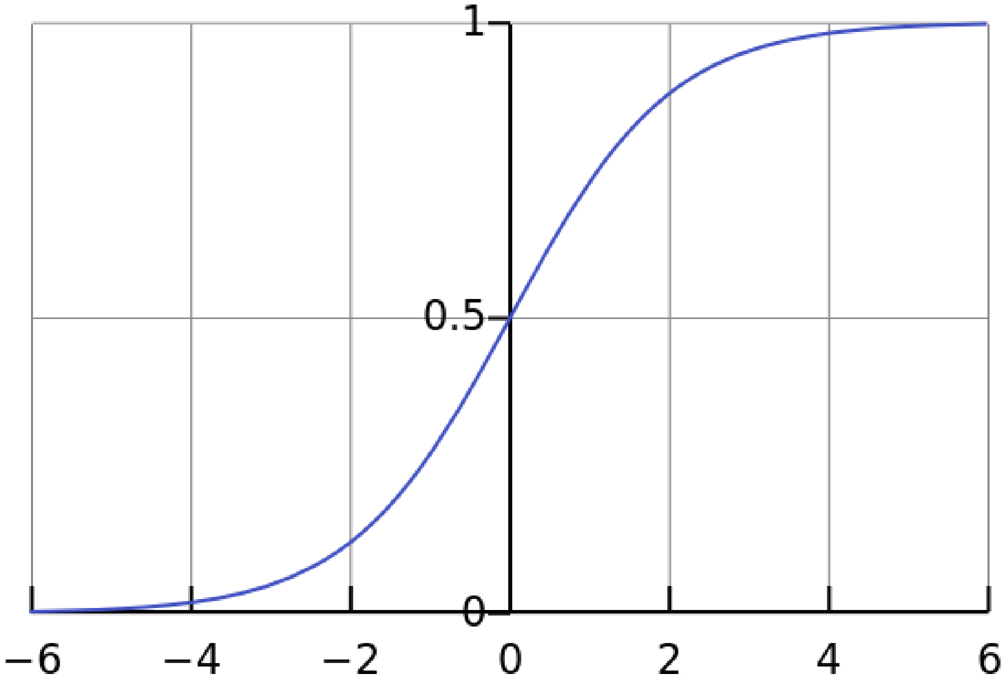

Figure 3.2: Logistic Activation function

Alternatively, the Hyperbolic Tangent (tanh) Function, which is also a Sigmoid and Squashing Function, can serve as an Activation Function.

$$\psi(u) = \frac{\exp(u) - \exp(-u)}{\exp(u) + \exp(-u)}, \quad -\infty < u < \infty$$

Compared to the Logistic Function, this function is a rescaled versions of the Logistic Sigmoid to assume negative values and is bounded between -1 *and* 1. It approaches $1(-1)$ when its argument goes to $\infty(-\infty)$. This function is more flexible because the negative values, in effect represents " suppressing" signals from the Hidden Units. It should be noted that for Logistic Function ψ, the re-scaled function $\hat{\psi}$ such that $\hat{\psi}(u) = 2\psi(u) - 1$ also generates values between -1 *and* 1 and may be used in place of the *tanh* function.

The advantage of the rescaled versions is that they shifts data inputs around zero because the magnitude of the derivatives is greater for these values which enables a faster training. In addition, using an activation function that outputs in the interval $[0, 1]$ makes big negative values of the input saturating to zero which has a negative effect on the training because they get stuck in the current state and consequently the computational time increases. The additional problem of having only positive values of the input is that, all of the weights that feed into a node can only increase or decrease all together in the training phase for a given input pattern which creates an inefficient update path.

The aforementioned Activation Functions are chosen for convenience because they are differentiable everywhere and their derivatives are easy to compute. In particular, when ψ is Logistic Function,

$$\frac{d\psi(u)}{du} = \psi(u)[1 - \psi(u)];$$

where ψ is the $tanh$ function,

$$\frac{d\psi(u)}{du} = \left[\frac{2}{\exp(u) + \exp(-u)}\right]^2 = sech^2(u)$$

These properties facilitates Parameter Estimation. The Activation Function in the Output Layer, it is common to set it as the Identity Function so that the Output enjoys the freedom of assuming any real values. When the target is binary variable taking values zero and one, as in the Classification Problems, the Activation Function in the Output Layer may be chosen as Logistic so that the Outputs must fall between zero and one, analogous to a Logit Model.

3.5 Finite Population Total Estimator Based on Feedforward Backpropagation Neural Network

In this section, the Estimator of Finite Population Total based on Feedforward Backpropagation Neural Network is discussed and presented.

Let

$$T = \sum_{i \in s} y_i + \sum_{i \in r} y_i \tag{3.32}$$

be the Finite Population Total where s are the Sample Units and r are the Nonsampled Units. Assume that y_i is given according to equation (1.4) with $x_i \in \mathbb{R}^d$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ *i.i.d.* Consider Estimating $m(x)$ based on approximating it by Feedforward Backpropagation Neural Network. As the basic building block, we consider the Neurons as a Nonlinear transformation of a Linear Combination of the Input $x = (x_1, \ldots, x_d)'$.

Feedforward Neural Network with more than one layer of Hidden Units are more complicated Network which allow feedback of information can be specified and considered but for simplicity, the study only dealt with the structure presented in equation (3.33), which is commonly used for a wide variety of applications and has the appealing feature of being implemented in statistical softwares In the simplest case of one Hidden Layer with $H \geq 1$ Neurons, the Network in equation (3.23) can be rewritten to represent the Network Function as follows

$$f_H(x,\theta) = v_0 + \sum_{h=1}^{H} v_h \psi\left(w_{0h} + x^T w_h\right), \; x \in \mathbb{R}^d \tag{3.33}$$

with $w_h = (w_{1h}, \ldots, w_{dh}) \in \mathbb{R}^d$ and

$$\begin{aligned}\theta = \big(w_{01}, \ldots, w_{0H}, w_1^T, \ldots, & \\ & w_H^T, v_0, \ldots, vh^T \in \mathbb{R}^{M(H)}, \\ & where \; M(H) = (d+1)H + H + 1\end{aligned} \tag{3.34}$$

represents the vector of all Parameters Weights of the Network. $\psi : \mathbb{R} \longmapsto \mathbb{R}$ is a given Activation Function. For Regression problems, function of Sigmoid shape, for instance, looking like the distribution function of a real random variable frequently provides good results. $f_H(x;\theta)$ specifies a mapping from the input space $\mathbb{R}^d$ to the output space which for this study is one dimensional. Such class of all Network output functions $O = \left\{f_H(x;\theta), \; \theta \in \mathbb{R}^{M(H)}, \; H \geq 1\right\}$ has several uniform approximation properties (Funahashi, 1989; Cybenko, 1989; White, 1990), e.g for any continuous function m, any $\varepsilon > 0$ and any compact set $C \subseteq \mathbb{R}^d$ there exist a function $f_H \in O$ with

$$\underset{x \in C}{sup} \mid m(x) - f_H(x;\theta) \mid < \epsilon$$

These implies that, any Regression Function $m(x)$ may be approximated arbitrary well using a large enough number of Neurons and appropriate Parameters θ.

Therefore, a Nonparametric Estimate for $m(x)$ is obtained if H is chosen first, which serves as a tuning Parameter and determines the smoothness of the Estimate and then Estimate the Parameter θ from the data by Nonlinear Least Squares

$$\hat{\theta}_n = arg \underset{\theta \in \Re^{M(H)}}{min} D_n(\theta) \tag{3.35}$$

with

$$D_n(\theta) = \sum_s (y_i - f_H(x;\theta))^2$$

Under appropriate conditions, $\hat{\theta}_n$ converges in Probability for $n \to \infty$ and a constant H to the Parameter vector $\theta \in \Theta_H$ which corresponds to the best approximation of $m(x)$ by a function of type $f_H(x;\theta),\ \theta \in \Theta_H$ with

$$\theta = arg \min_{\theta \in \Re^{M(H)}} D(\theta) \;\; with \;\; D(\theta) = E\left\{m(x) - f_H(x;\theta)\right\}$$

Also, under some stronger assumptions, the Asymptotic Normality of $\hat{\theta}_n$ and thus the Estimator of $\hat{m}(x) = f_H(x;\hat{\theta}_n)$ also follows for the Regression Function $m(x)$, (Franke and Neumann, 2000). Therefore, the immediate consequence of these is that $f_H(x;\hat{\theta}_n) \longrightarrow f_H(x;\theta)$ for $n \longrightarrow \infty$, (White, 1990)
The Estimation Error $\hat{\theta}_n - \theta$ can be divided into two Asymptotically independent sub-components : $\hat{\theta}_n - \theta = (\hat{\theta}_n - \theta_n) + (\theta_n - \theta)$, where the value

$$\theta_n = arg \min_{\theta \in \Re^{M(H)}} \sum_{i=1}^{n} \left\{m(x) - f_H(x,\theta)\right\}^2$$

minimizes the Sample version of $D(\theta)$, (Franke and Neumann, 2000). By Universal Approximation property of Neural Network, $f_H(x;\theta)$ converges to the Regression Function $m(x)$ for $H \longrightarrow \infty$. Therefore $f_H(x;\hat{\theta}_n)$ should become a consistent Nonparametric Estimate of $m(x)$ if H increases with n and with an appropriate rate. From these results, the corresponding estimate of the Finite Population Total is therefore, given as

$$\hat{T}_{NN} = \sum_{j \in s} y_j + \sum_{j \in r} \hat{m}_n(x_j) \tag{3.36}$$

where $\hat{m}_n(x_j) = f_H(x;\hat{\theta}_n)$
Thus equation (3.36) is the developed Estimator for the Finite Population Total.

The following comments are made about this Estimator. First, $\hat{T}_{NN}$ is a Model-Based Estimator, so that all the inference is with respect to the Model for the $y_i^{'}s$, not the Survey Design. Next, this Estimator is identical to that proposed in (Dorfman, 1992a), except that the NN is replaced by a Kernel-Based Regression. Lastly, this Estimator can be used to Estimate the Population Total of a Finite Population as long as each of the Nonsampled elements has the same distribution as the Sample.

3.5.1 Regularity Notes on the Proposed Estimator

For fixed H we just fit a Nonlinear Regression Model to the data. However, we are aware that this model will be misspecified and that we have to select a decent H, determining the form of the Nonlinear Regression Function and the dimension of its Parameter, to get a reasonable balance between bias and variance of $\hat{m}_n(x)$ as an Estimate of $m(x)$.

1. The Parameter vector θ of equation (3.33) is not uniquely determined by the function, i.e for different values of θ we get the same function $f_H(x, \theta)$. If, for example the Activation Function is Anti-symmetric, $\psi(-x) = -\psi(x)$, like the Logistic Function in equation (3.31), then changing the enumeration of Hidden Units and multiplying all Weights $w_{ih},\ i = 1, 2, \ldots, d$, going into a Hidden Units and simultaneously the Weight v_h going out of the Neuron by -1 do not change the function.

2. To avoid this ambiguity and the obvious problems with Estimation we consider only Parameter Vectors in a subset $\Theta_H \subset \mathbb{R}^{M(H)}$ chosen such that to each function in equation 3.33 with H Neurons, their exists exactly one corresponding Parameter Θ_H. For Anti-symmetric ψ we can choose for example $\Theta_H = \left\{\theta \in \mathbb{R}^{M(H)};\ v_1 \geq v_2 \geq \ldots \geq v_H\right\}$, that is, the last h coordinates of θ are in decreasing order. For more details on the identification of Parameters see (Hwang and Ding, 1997).

For appropriate choice of $\theta,\ f_H(x, \theta)$ will approximate a Linear Function such that the Estimator reverts to classical regression Estimator. If for example $H = 1$ we have $f_{x;\theta} = v_0 + v_1\psi\left(w_{01} + x^T w_1\right)$. Choosing the Logistic ψ of equation 3.31 and letting $\|w_h\| \longmapsto 0$ the a Taylor expansions of ψ up to order 1 gives

$$f_H(x, \theta) = v_0 - v_1 + \frac{2v_1}{1 + c_0}\left\{1 + \frac{2c_0}{1 + c_0} x^T w_1\right\} + o\left(w_1\right)$$

with $c_0 = \exp(-2w_{01})$, which is an approximately linear function.

Theoretically, Feedforward Neuron Network with one hidden layer suffice by the universal approximation property. In practice, Networks with more than one Hidden Layer may provide a better approximation to $m(x)$ with fewer Parameters, see (Cybenko, 1989; Funahashi, 1989; Barron, 1994; Montanari and Ranalli, 2003; Asnaashari et al., 2013).

In the next section Theoretical properties of the proposed Estimator are considered.

3.6 Theoretical Properties of the Proposed Estimator

3.6.1 Assumptions

To be able to prove the theoretical properties, the following assumptions are made;

A1) The errors ε_i are $i.i.d$ with mean 0, Finite Variance σ^2 satisfying

$$pr\left(|\varepsilon_i| > t\right) \leq a_0 \exp\left\{-a_1 t^{\alpha}\right\} \text{for all } t \geq 0$$

and for some $a_0,\ a_1$ and $\alpha > 0$.

A2) The Auxiliary measurements $x_i \in \mathbb{R}^d$ are $i.i.d$ with an absolutely continuous distribution F having a Finite Second Moment.

$$\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_d} f(t_1,\ldots,t_d)dt \tag{3.37}$$

where $f(.)$ is strictly positive density whose support is a compact subset of $\mathbb{R}^d$. Moreover,

$$pr\left(\| x_i \| > t\right) \leq b_0 \exp\left\{-b_1 t^{\beta}\right\} \text{ for all } t \geq 0 \tag{3.38}$$

and for some $b_0,\ b_1$ and $\beta > 0$. A3) $m(x)$ is a bounded function.

A4) For each sequence of Finite Population indexed by v, conditioned on the value x_i, the Superpopulation Model equation (1.4), where ε_i satisfies $A1$, then, the x_i are considered fixed with respect to the Superpopulation Model ξ.

A5) The Survey variable has a bounded moment with $\xi-$probability 1. Moreover, its noted that $(A1),\ldots,(A3)$ immediately imply for some $c_0, c_1 > 0$

$$Pr(|y_i| > t) \leq c_0 \exp\{-c_1 t^{\alpha}\}, \text{ for all } t \geq 0 \tag{3.39}$$

A6) The Sampling rate is bounded, that is

$$\lim_{v \longrightarrow \infty} sup \frac{n}{N} = \pi, \text{where } \pi \in (0,1)$$

A7) The Parameter space Θ is a compact set, θ an interior point of Θ and it is irreducible; that is for $h,\ h^{'} \neq 0$ none of the following three cases holds Hwang and Ding (1997).

a) $v_h = 0,$ for some $h = 1, \ldots, H$

b) $w_h = 0,$ for some $h = 1, \ldots, H$

c) $(w_h^{'}, w_{0h}) = \pm(w_{h'}^{'}, w_{0h'}),\ for\ w \neq w^{'}$

A8)The Activation Function ψ in equation (3.30) is Asymmetric Sigmoid Function differentiable to any order. Additionally, we make an assumption that the class of functions $\{\psi(b_t, b_0), b > 0\} \cup \{\psi \equiv 1\}$ is linearly independent. The Logistic Activation Function in equation 3.31 fulfills these requirement.

Therefore, to prove Consistency of $\hat{T}_{NN}$, the rate which determines how the complexity of the Network and the possible roughness of the Function Estimate $\hat{m}_n(x)$ increases with the Sample size n has to satisfy some conditions. We follow (White, 1990) and restrict the number H of Neurons and the overall size of the Network Weights v_h, w_{jh} simultaneously. For some sequences $H_n, \Delta_n \longrightarrow \infty$, let

$$\Theta_n = \Theta\left(H_n, \Delta_n\right) = \left\{\theta \in \Theta; \sum_{h=0}^{H_n} |v_h| \leq \Delta_n, \sum_{h=1}^{H_n}\sum_{j=0}^{d} |\omega_{hj}| \leq H_n\Delta_n\right\} \tag{3.40}$$

For given sample size n, we consider only Network Function in

$$O_n = O\left(H_n, \Delta_n\right) = \{f_{H_n}(x,\theta); \theta \in \Theta(H_n, \Delta_n)\} \tag{3.41}$$

as an Estimate for $m(x)$. Therefore, we redefine the Parameter Estimate as

$$\hat{\theta}_n = \underset{\theta \in \Theta_n}{argmin} \sum_{s} \left(y_i - f_H(x; \theta)\right)^2 \tag{3.42}$$

and the Network Estimate for $m(x)$ is therefore given by

$$\hat{m}_n(x) = f_{H_n}(x, \hat{\theta}_n) \tag{3.43}$$

which is a kind of a Sieve Estimate in the sense of (Grenander and Ulf, 1981; Geman and Hwang, 1982).

To prove for Consistency of $\hat{T}_{NN}$, first we need to show that the Neural Network based Regression Function $\hat{m}_{NN}$ is normally distributed and also Consistent.

Theorem 3.6.1 *(Franke and Neumann, 2000): Suppose that certain Conditions a are satisfied and for $n \to \infty$,*

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_n - \theta_n \\ \theta_n - \theta \end{pmatrix} \xrightarrow[d]{} N\left(0, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right)$$

that is $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ *and* $\sqrt{n}\left(\theta_n - \theta\right)$ are Asymptotically Independent Normal Random Vectors with Covariance Matrices Σ_1 *and* Σ_2, respectively, where

$$\Sigma_1 = A(\theta)^{-1} B_1(\theta) A(\theta)^{-1}, \quad \Sigma_2 = A(\theta)^{-1} B_2(\theta) A(\theta)^{-1}$$

with

$$B_1(\theta) = 4. \int \sigma_\epsilon^2(x) \nabla f_H(x,\theta). \nabla f_H(x,\theta)' F(x) dx$$

$$B_2(\theta) = 4. \int (m(x) - f_H(x,\theta))^2 \nabla f_H(x,\theta). \nabla f_H(x,\theta)' F(x) dx$$

and $A(\theta) = \nabla^2 D(\theta)$

As noted by (Franke and Neumann, 2000), as an immediate consequence, $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$ is Asymptotically Normal with Mean 0 and Covariance Matrix $\Sigma_1 + \Sigma_2$. In the correctly specified case where $m(x) = f_H(x, \theta_0)$, the Σ_2 is equal to the zero matrix, as there is no effect due to the randomness of the $X_i's$, that is $\theta_n = \theta$. In the Misspecified case, the randomness of the inputs causes a difference of order $n^{-\frac{1}{2}}$ between the optimal Parameters θ_n and θ. A simple complete proof of these theorem is given in Theorem 1 of (Franke and Neumann, 2000) and also see the results of Theorem 5.1 of (Shen et al., 2019).

From this Theorem, it follows that, the Neural Network Estimator of the Mean Function $m(x) = f_H(x, \theta_0)$ is Asymptotically Normal, then it also follows that, equation (3.36) which is the developed Estimator of the Finite Population Total base on the Feedforward Backpropagation Neural Network is also Asymptotically Normal.

Theorem 3.6.2 *(Franke and Diagne, 2006): Let* $(y_1, x_1), \ldots, (y_n, x_n)$ *be i.i.d variable with* $y_i \in \mathbb{R}$, *and* $x_i \in \mathbb{R}^d$. *Let the distributions of* y_i *and* x_{x_i} *satisfy A2 and equation* 3.37. *Let* $O_n = O(H_n, \Delta_n)$, $n \geq 1$ *be the set of Neural Network Functions given by equation (3.41) with an Activation Function* ψ *which is Lipschitz continuous on* $\mathbb{R}$, *strictly increasing and satisfying equation (3.30). Let* $\hat{m}_n(x) = E(y_i|x_i) = x$ *be in the closure of* $\bigcup_{n=1}^{\infty} O_n$ *in* $L^2(F)$ *that is, in the space of functions Square Integrable with respect to the distribution of* x_i.
Then $\hat{m}_n(x)$ *is a Consistent Estimate of* $m(x)$ *in the* $L^2(F)$*-sense , that is*

$$\int (m(x) - \hat{m}_n(x))^2 \, dF(x) \longrightarrow 0 \textit{ in probability} \tag{3.44}$$

provided that $H_n, \Delta_n \longrightarrow \infty$ *such that* $\Delta_n = o(n^{\frac{1}{4}})$
$H_n, \Delta_n^4 log\ n = o(n)$ *and* $H_n log\ n = o(\Delta_n^{\alpha})$
where α *determine the rate of decrease of the tail of the distribution of the* y_i *by equation (3.39)*

3.7 Proof

Here,the highlight of the proof of Theorem 3.6.2 is provided. The theorem can be proven exactly as Theorem 2.1 of (Franke and Diagne, 2006) for stationary processes satisfying an α-mixing condition and also as Theorem 3.1 of (Shen et al., 2019) for fixed data. As here the data are independent, the Bernstein Inequality for stationary processes may be replaced by a Bernstein Inequality for independent data like that one in section (2.5.4) of Lemma A of (Serfling, 2000, 2009). Therefore, the right hand side of equation (5.1) of (Franke and Diagne, 2006) changes to

$$c_1 \exp\left(-c_2 \frac{\Delta}{N M_N^2}\right) \text{ instead of } c_1 \exp\left(-c_2 \frac{\Delta^2}{\sqrt{N} M_N^2}\right)$$

Then the proof proceeds exactly as in (White, 1990).

It should be noted that, for bounded Random Variables (y_i, x_i), the last condition on H_n, Δ_n involving α can be dropped. In that case, Theorem 3.6.2 essentially is equivalent to Theorem 3.3 of (White, 1990). The Parameters $H_n,\ \Delta_n$ which determines the Network Complexity and therefore the Smoothness of the Function Estimate can be determined adaptively from the data by Cross Validation without changing the Consistency of $\hat{m}_n(x)$ using Theorem 3.4 of (White, 1990). For the detail and complete proof of these theorem, see the work of (White, 1990; Franke and Diagne, 2006; Shen et al., 2019).

This Theorem is important in the development of this work in sense that, showing that the developed estimator of the Finite Population Total given in equation (3.36) is Consistent, it is required that the Neural Network estimator of the Mean Function is also Consistent.

Equation (3.44) is in the integral form thus using it the proving of the foregoing work of Consistence Population Total in equation (3.36) might be a little complex. Therefore, equation (3.44) is required with a simple mean over the unobserved $x_i,\ i \in r$ instead of the integral. to allow as to prove the consistency of $\hat{T}_{NN}$. The following results shows that the difference between the integral and the Simple mean is negligible.

Theorem 3.7.1 *Let $((y_1, x_1), \ldots, (y_N, x_N))$ be i.i.d with equation (1.4) for some bounded $m(x)$. Let F denote the distribution of x_i. Let $|\psi(u)| \leq 1$, and $s = 1, \ldots, n$ be the index set of the observed data and $r = n+1, \ldots, N$ the index of unobserved data. Let $\hat{\theta}_n$ be defined as in equation* 3.35 *with $\hat{m}_n(x) = f_{H_n}(x, \hat{\theta}_n)$ denote the Estimate of $m(x)$ based on the Sample $(y_i, x_i),\ i \in s$. Let $n, N \longrightarrow \infty$ such that $\frac{n}{N} \longrightarrow \pi(0,1)$ and let H_n, Δ_n satisfy conditions in Theorem (3.7.2). Then for $\delta > 0$*

$$\begin{aligned} &Pr\left(\left|\frac{1}{N-n}\sum_{j\in r}\right.\right. \\ &(m(x_j) - \hat{m}_n(x_j))^2 - \int (m(x_j) - \hat{m}_n(x_j))^2 dF(x) > \delta|(y_i, x_i),\ i \in s \\ &\leq d_1 \exp\left\{-d_2 \frac{N\delta^2}{\Delta_n^4}\right\} \end{aligned} \tag{3.45}$$

for all $\delta > 0$ and all N large enough where $d_1,\ d_2$ are some constants independent of N, n *and* $(y_i, x_i),\ i \in s$

3.7.1 Proof

From assumption A3, let C be the upper bound of $m(x)$. By definition of $\hat{m}_n(x) = f_{H_n}(x, \hat{\theta}_n)$ and $\hat{m}_n(x) \in O(H_n, \Delta_n)$, we immediately have

$$|\hat{m}_n(x)| \leq \Delta_n \text{ a.s } \quad |\psi(u) \leq 1|$$

if we set

$$V_{N_i} = (m(x_j) - \hat{m}_n(x_j))^2 - \int (m(x_j) - \hat{m}_n(x_j))^2 dF(x), \ i \longrightarrow r \tag{3.46}$$

these therefore results to

$$\begin{aligned} &|V_{N_i}| \leq 4(C^2 + \Delta_n^2) \\ &E\left\{V_{N_i}|(y_i, x_i), \ i \in s\right\} = 0 \\ &E\left\{V_{N_i}^2|(y_i, x_i), \ i \in s\right\} \leq 32(C^4 + \Delta_n^4) \end{aligned} \tag{3.47}$$

note that $\hat{m}_n(x)$ is independent of $(y_i, x_i), i \in r$, and completely determined by $(y_i, x_i), i \in s$. Now applying Bernsteins inequality (Lemma A, section 2.5.4) of (Serfling, 2000) in equation (3.46), we get

$$\begin{aligned} &Pr\left(\frac{1}{N-n}\left|\sum_{j\in r} V_{N_j}\right| > \delta(y_i, x_i), i \in s\right) \\ &\leq 2\exp\left\{-\frac{(N_n)\delta^2}{64(C^4 + \Delta_n^4) + \frac{2}{3}4(C^2 + \Delta_n^2)\delta}\right\} \end{aligned} \tag{3.48}$$

Now the results follow as $\Delta_n \longrightarrow \infty$ and therefore Δ_n^4 dominates the denominator of the exponent for N large enough and as $N - n$ coincides Asymptotically with $(1 - \pi)N$. Moreover, as $\Delta_n = o(n^{\frac{1}{4}})$, $\frac{N}{\Delta_n^4} \longrightarrow \infty$, that is, the right hand side of the inequality converges to zero(taking limits as $\Delta_n \longrightarrow \infty$).

3.7.2 Consistency of the Developed Estimator of Finite Population Total

Theorem 3.7.2 *If (A1)-(A8) are satisfied and if the Activation Function $\psi(u)$ is Lipschits continuous and strictly increasing and satisfies equation (3.30) also Theorem 3.6.2 holds, then the Neural Network Estimate $\hat{T}_{NN}$ of the Population Total T given by equation (3.36) with $\hat{m}_n(x) = f(x, \hat{\theta}_n)$ and $\hat{\theta}_n$ given by equation*

(3.35) is Consistent in the following sense.

$$\begin{aligned}
&\frac{1}{N}\left|T-\hat{T}_{NN}\right| \longrightarrow 0 \text{ in probability}\\
&\text{where N, n} \longrightarrow \infty \text{ with } \frac{n}{N} \longrightarrow \pi \in (0,1)
\end{aligned} \tag{3.49}$$

provided that the number H_n and the bound Δ_n of the Network Weights satisfy $H_n, \Delta_n \longrightarrow \infty$ such that

$$\begin{aligned}
&\Delta_n = o(n^{\frac{1}{4}})\\
&H_n\Delta_n^4 \log n = o(n)\\
&H_n \log n = o(\Delta_n^{\alpha})
\end{aligned} \tag{3.50}$$

where α determines (by A1) how fast the tail probability of the ε_i and y_i decreases.

White (1990) showed that, the appropriate choice for Δ_n is such that $\Delta_n \longrightarrow \infty$ as $n \longrightarrow \infty$ and $\Delta_n = o(n^{\frac{1}{4}})$, i.e $n^{\frac{1}{4}}\Delta_n \longrightarrow 0$ as $n \longrightarrow \infty$

Proof

$$\begin{aligned}
\frac{1}{N}\left|T-\hat{T}_{NN}\right| &= \frac{1}{N}\left|\sum_{j\in r}(y_j - \hat{m}_n(x_j))\right|\\
&= \frac{1}{N}\left|(m(x_j) - \hat{m}_n(x_j)) + \sum_{j\in r}\varepsilon_j\right|\\
&\leq \frac{1}{N}\left|(m(x_j) - \hat{m}_n(x_j))\right| + \frac{N-n}{N}\frac{1}{N-n}\left|\sum_{j\in r}\varepsilon_j\right|\\
&\leq \frac{1}{N}(m(x_j) - \hat{m}_n(x_j))^2 + \frac{N-n}{N}\frac{1}{N-n}\left|\sum_{j\in r}\varepsilon_j\right|
\end{aligned} \tag{3.51}$$

by Jensen's inequality.
Now the last term converges to

$$\frac{N-n}{N}\frac{1}{N-n}\left|\sum_{j\in r}\varepsilon_j\right| = (1-\pi)\left|E(\varepsilon_j)\right|$$

where $(1-\pi)|E(\varepsilon_j)| = 0$ since $E(\varepsilon_j) = 0$ by law of large numbers. The first term of equation (3.51) decomposes into

$$\begin{aligned}&\frac{N-n}{N}\left(\frac{1}{N-n}\sum_{j\in r}(m(x_j)-\hat{m}_n(x_j))^2-\int(m(x_j)-\hat{m}_n(x_j))^2dF(x)\right)\\&+\frac{N-n}{N}\int(m(x_j)-\hat{m}_n(x_j))^2dF(x)\end{aligned} \tag{3.52}$$

The right hand terms of equation (3.52) converges to 0 by Theorem 3.6.2 and as $\frac{N-n}{N}\longrightarrow 1-\pi$.

The proof is completed by using Theorem 3.7.1 to cope with left hand terms where we drop the factor $\frac{N-n}{N}$ converges to $1-\pi$ any how.

$$\begin{aligned}&Pr\left(\left|\frac{1}{N-n}\sum_{j\in r}(m(x_j)-\hat{m}_n(x_j))^2-\int(m(x_j)-\hat{m}_n(x_j))^2dF(x)\right|>\delta\right)\\&=E\left\{Pr\left[\left|\frac{1}{N-n}\sum_{j\in r}\left(m(x_j)-\hat{m}_n(x_j)\right)^2-\int(m(x_j)-\hat{m}_n(x_j))^2dF(x)\right|\right.\right.\\&>\delta|(y_i,x_i,\ i\in s)\\&\leq d_1\exp\left\{-d_2\frac{N\delta}{\Delta_n^4}\right\}\longrightarrow 0,\quad\forall\ \delta\\&>0,\ \ \Delta_n\longrightarrow\infty,\quad n\longrightarrow\infty\end{aligned} \tag{3.53}$$

hence the proof.

3.7.3 Mean Square Error(MSE) of the Developed Estimator

Mean Square Error is used to measure the accuracy of the Estimator among other measures of performance. The MSE is define by $E\left(\hat{T}_{NN}-T\right)^2$ where T denotes the true Population Total. To Estimate $E\left(\hat{T}_{NN}-T\right)^2$, first, we consider

$$E\left[\left(\hat{T}_{NN}-T\right)^2\right]=E\left[\left(\sum_{i=1}^{H}\sum_{j=n+1}^{N}\hat{m}(x,\theta)-\sum_{j=n+1}^{N}\left(m(x)+\epsilon\right)\right)^2\right]$$

$$=\frac{(N-n)^2}{N^2}E\left[\left(\frac{1}{N-n}\sum_{i=1}^{H}\sum_{j=n+1}^{N}\hat{m}(x)-\sum_{j=n+1}^{N}\left(m(x)+\epsilon\right)\right)^2\right]+\frac{N-n}{N}var(\epsilon_i)$$

$$= \frac{(N-n)^2}{N^2}$$

$$E\left[\left(\frac{1}{H(N-n)}\sum_{i=1}^{H}\sum_{j=n+1}^{N}\hat{m}(x,\theta) - E\left(T_k|D,X_j\right) + E\left(T_k|D,X_j\right) - E\left(T_k\right)\right)^2\right]$$

$$+ \frac{N-n}{N}var(\epsilon_i)$$

$$= \frac{\tau_D^2}{H}(1-f)\left\{E\left(T_k|D,X_j\right) - E\left(T_k\right)\right\}^2 + \frac{1-f}{N}var(\epsilon_i) \qquad (3.54)$$

where the $X_j = (x_{n+1},\ldots,x_N)$ are a set of Nonsampled Auxiliary Units. T_k denotes the Total of the Nonsampled elements and $E(T_k) = \sum_{j=n+1}^{N} m(x)$.
The last approximation of equation (3.54) follows from equation 15 of (Liang and Kuk, 2004), that is

$$E\left(\frac{1}{HN}\sum_{i=1}^{H}\sum_{j=n+1}^{N}\hat{m}(x,\theta) - (1-f)E(T_k)|D,X_j\right)^2 \approx \frac{\tau_D^2}{H}$$

for some positive constant τ_D^2.

The term $E\left(T_k|D,X_j\right) - E\left(T_k\right)$ is the predictor bias due to randomness or Sampling Bias of D. Now from equation (3.54), we have

$$E\left(T_{NN} - T\right)^2 = E\left(\frac{\hat{\tau}_D^2}{H}\right) + (1-f)^2 E\left\{E\left(T_k|D,X_j - \right.\right.$$
$$E\left(T_k\right)^2 +$$
$$\frac{1-f}{N}var(\epsilon_i)$$

(3.55)

As noted in (Liang and Kuk, 2004), the quantity τ_D^2 can be Estimated by batch method. Therefore,

$$\hat{\tau}_D^2 = \frac{s}{r-1}\sum_{t=1}^{r}\left(\hat{T}_{NN,t} - T_{NN}\right)^2 \qquad (3.56)$$

for details see (Liang and Kuk, 2004). Equation 3.56 can be substituted in equation (3.55) in lieu of $E(\tau_D^2)$.
Now, under the assumption that the $\frac{\epsilon_i}{\sigma} \sim t(v)$, then the Estimate of $var(\epsilon_i)$ is

given as

$$v\hat{a}r(\epsilon_i) = \frac{v}{v-2}\frac{1}{H}\sum_{i=1}^{H}\hat{\sigma}_i^2 \tag{3.57}$$

Under the assumption that the Population is made up of exact copies of the Sampled (training) data, we have $E\left(T_k|D, X_j\right) - E\left(T_k\right) \approxeq \hat{T} - T$ where $\hat{T}$ the fitted Sample Total and

$$E\left(\hat{T} - T\right)^2 = \left(\sum_{i=1}^{n}\hat{\epsilon}_i\right)^2 = Var(\hat{\epsilon}_i) \tag{3.58}$$

Under the true model, we have $Var(\hat{\epsilon}_i) = var(\epsilon_i)$. Hence the $E\left\{E\left(T_k|D, X_j - E\left(T_k\right)\right)\right\}^2$ can be Estimated by

$$B\hat{i}as^2 = \frac{1}{n}v\hat{a}r(\epsilon_i) \tag{3.59}$$

Thus, $E\left(T_{NN} - T\right)^2$ can be Estimated by

$$\begin{aligned}\hat{E}\left(T_{NN} - T\right)^2 &= \frac{\hat{\tau}_D^2}{H} + (1-f)B\hat{i}as^2 + \frac{1-f}{N}tvar(\epsilon_i)\\ &= \frac{\hat{\tau}_D^2}{H} + \frac{1-f}{n}v\hat{a}r(\epsilon_i)\end{aligned} \tag{3.60}$$

As $H \longrightarrow \infty$, equation (3.60) reduces to

$$\hat{E}\left(T_{NN} - T\right)^2 = \frac{1-f}{n}v\hat{a}r(\epsilon_i) \tag{3.61}$$

3.8 Comparison of the Developed Feedforward Backpropagation Neural Network Estimator with GAM, MARS and LP Estimators

In this section a description of methodology used to compare the developed Nonparametric Feedforward Backpropagation Neural Network Estimator

$$\hat{T}_{NN} = \sum_{j\in s} y_j + \sum_{j\in r}\hat{m}_n(x_j) \tag{3.62}$$

where $\hat{m}_n(x_j) = f_H(x; \hat{\theta}_n)$
with Generalized Additive Model estimator define in equation 3.22

$$\hat{T}_{GAM} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{m}(x_i)_{GAM} \tag{3.63}$$

Multivariate Additive Model Estimator define in equation 3.19

$$\hat{T}_{MARS} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{m}(x_i)_{MARS} \tag{3.64}$$

and Local Polynomial Estimator

$$\hat{T}_{MARS} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{m}(x_i)_{LP} \tag{3.65}$$

is given. Details of these functions are in sections 2.2.1 and 3.4. This comparison is through a simulation study in which some high dimension Superpulation models were first considered, from which data was generated using Monte-Carlo methods. The models used in this Thesis for simulation are adopted from the work done by (Feng and Simon, 2017), three scenarios where the true function is the sum of two dimensional linear function, two dimensional quadratic function and a three dimensional mixed function is considered.

From the generated dataset, a sample was taken using Simple Random Sampling without replacement and used as a training set to calibrate the Neural Network. The calibrated Neural Network was used to estimate the nonsampled units for the variable of interest Y. Various replications were conducted, and for each iteration the Bias, MSE and Mean Absolute Errors were noted.This was done for various estimators and the respective results finally averaged. The estimator with the smallest Bias, MSE and Mean Absolute Errors was considered to be the best estimator of the Finite Population Total in high dimension.

3.9 Determination of the Smoothing Parameters

The Estimate of the Bandwidth h for the Estimator based on Local Polynomial was obtained by determining the Smoothing Parameter that Minimizes the Least Squares Crossvalidation for the Finite Population Total Estimator for the characteristic of the variable under consideration. This method applies to the Local Polynomial Regression Estimator. This is due to the fact that the procedures of these Estimator incorporates the Smoothing Procedures in their estimation. Therefore, the value of the Smoothing Parameter is necessary in the construction of the appropriate Smoothers in order to Estimate the Population Total. The search for the Smoothing Paramters is constrained in the range $\frac{1}{4}n^{\frac{-1}{5}}\sigma < h < \frac{3}{2}n^{\frac{-1}{5}}$(Silverman, 1986) and $|k| \leq 1$ where $k = \frac{x_i - x_j}{h}$

3.10 Performance Criterion of Estimators

The Estimates of the Finite Population Total for the Neural Network, Multivariate Additive Regression spline, Generalized Additive models and Local Polynomial Regression Estimators are recorded analyzed and deductions made. The unconditional results for the Estimators were computed that are used in the analysis that acts as performance indicators of the Estimators. The results include; Bias,Mean Square Error(MSE), Mean Absolute Error(MAE) and Mean Absolute Percentage Error($MAPE$) respectively. These criterion are defined as follows;

i. Mean Square Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(T_i - \hat{T}_i\right)^2$$

ii. Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

iii. Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{|Y_i|} \times 100\%$$

CHAPTER FOUR
RESULTS AND DISCUSSIONS

4.1 Introduction

The problem being addressed in this thesis falls into the area of Survey Sampling as follows; there exists a parent population from which the variable of interested is associated with a number of Auxilliary Variables in which an appropriate Sampling technique is used to select a Sample that is used in the Estimation. Since the dateset is of high dimensional case, a Robust Estimator based on Feedforward Backpropagation Neural Network is developed and used to Estimate the Finite Population Total. The developed Estimator is then used on the data obtained from the United Nations Development Programme 2020 report to test its Robustness in real life application.

Thus, in this chapter, the theory developed in the previous chapters are tested here in a fairly wide range of sets of data. From a practical point of view, it is natural to inquire about the Finite Sample properties of the new Estimator of Finite Population Total based on Feedforward Backpropagation Neural Network and to compare it to popular Estimators for Population Total available in the literature.

The Estimation of the Population Total was done using five sets of data that include Simulated data and Secondary data obtained from the United Nations Development Programme 2020 report.

In order to understand how the Estimator developed in this thesis compared against other existing Nonparametric Regression Estimators, a comparison of the performance of the developed Estimator to that of identified Estimators based on Multivariate Adaptive regression Splines(MARS), Generalized Additive Models(GAM) and Local polynomial (LP) which can handle high dimensional data was performed. The performance measure used included the Bias, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

4.2 Description of the Population Data Sets

4.2.1 Simulation Settings

The Superpopulation Model in equation (1.4), used in this thesis follows from the problem proposed by (Feng and Simon, 2017). According to (Feng and Simon, 2017), three scenarios where the true function is the sum of Two dimensional Linear Function, Two dimensional Quadratic Function and a Three dimensional Mixed Function is considered. This models were considered in this thesis.

For all of the Simulation performed, data is generated according to model equation (1.4) where $\epsilon \sim N(0,1)$. The Auxiliary variable vector $X \in \mathbb{R}^d$ were generated from *iid* uniform(0,1) Random Vector. The Errors ϵ were generated from *iid* $N(0,1)$ with noise level $\sigma = 0.1, 0.4$. The *tanh* was used as the Activation Function for the Neural Network.

1,000 Samples of size 4,000 and 8,000, were generated using Simple Random Sampling from a Population of size 10,000. Because of the hypothesized relationship between the study variable and the auxiliary variable, which must be depicted in the Simulation, the Sampling is done with indices.

T_{NN}, with predictions obtained by means of the R function nnet() and by setting the number of units in the Hidden layer and the weight decay parameter as follows: (3,0.05), (6,0.15), (12,0.15), (6,0.2), (12,0.2); the weight decay Parameter is analogous to ridge regression introduced for Linear Models as a Solution to Collinearity. Larger values of it tend to favor approximations corresponding to small values of the Parameters of the net and therefore shrink the weights towards zero to avoid overfitting.

T_{GAM}, with predictions obtained through an additive splines model by means of the R function gam(). The number of degrees of freedom for the splines have been set equal to the values: $2, 3, 4, 5$

T_{MARS}, with predictions computed by means of MARS 3.6, the original collection of Fortran subroutines developed by (Friedman, 1991). The maximum interaction level has been fixed to 1 and with the following values of the number of basis functions: 5, 10, 15, 20.

T_{LP} Local Polynomial Regression Estimator with degree P=1 and bandwidth h=0.1 and h=0.25.

4.2.2 Two Dimensional Linear Model

$$m(x) = -1 + 2X_1 + 4X_2$$

Two Dimensional Quadratic Model

$$m(x) = 5.5 - 6X_1 + 8(X_1 - 0.5)^2 - 3X_2 + 32(X_3 - 0.5)^3$$

Three Dimensional Mixed Model

$$m(x) = 8(X_1 - 0.5)^2 + \exp(2X_2 - 1) + \sin(2\pi(X_3 - 0.5))$$

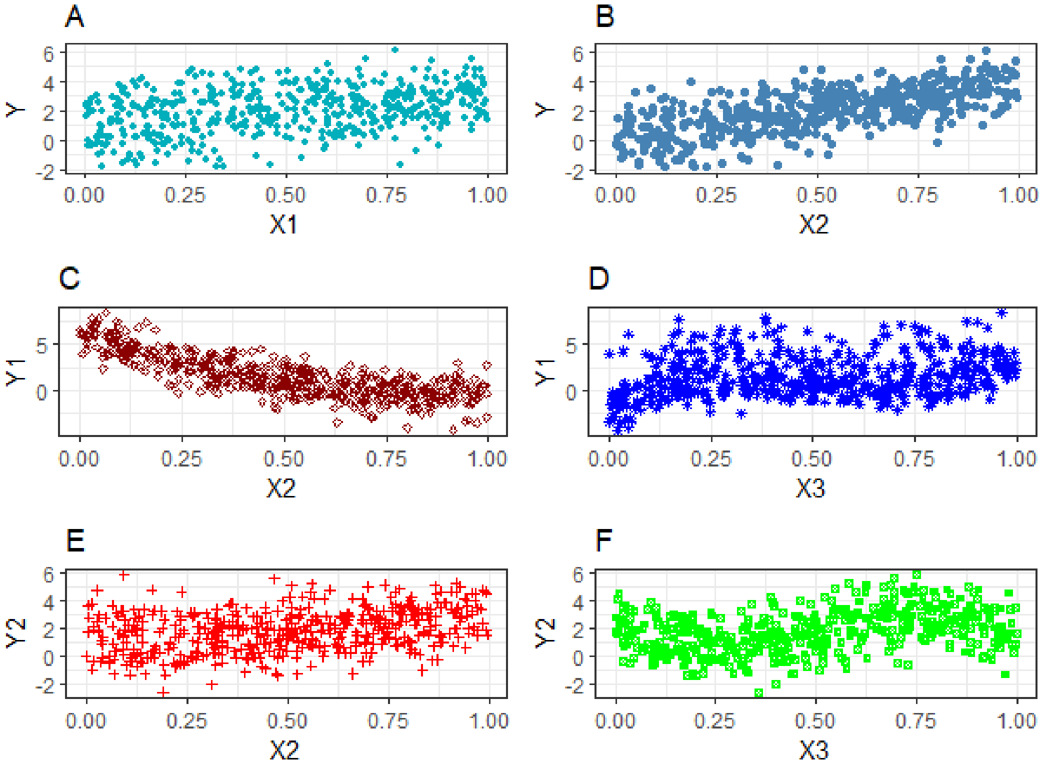

Figure 4.1: Relationship Between Y and X for a Two dimensional Linear Model,Two dimensional Quadratic Model and Three dimensional Mixed Model

From the scatter plots in Figure 4.1, shows various relationships that exists the response variable and the Auxiliary variables. In most of them, the relationships exhibited are Linear and Quadratic.

Tables 4.1-4.3 summarize the findings of this Simulation investigation. Unconditional Bias (UB), Unconditional Mean Square Error (UMSE), Unconditional Relative Mean Square Error (URMSE), and unconditional Mean Absolute Error (UMAE) for said Estimators at different Sample sizes are shown in Tables 4.1-4.3. The MAE reveals how near the Estimate being examined is to the true value, while the MSE and RMSE represent the Estimator's precision. For example, if TNN's UMSE and URMSE are comparable, it will reasonably be considered "better" or " more desirable" than other Estimators.

Table 4.1: Unconditional Bias, Mean Square Error, Relative root mean Square Error, Mean Absolute Error and Mean Absolute Percentage Error for Two Dimensional Linear Model

		Bias	MSE	RRMSE	MAE	MAPE
n=4000	$\hat{T}_{NN}$	8.7982	151.4639	0.1900	0.0043	0.4311
	$\hat{T}_{MARS}$	9.8620	153.1423	0.2170	0.0048	0.4785
	$\hat{T}_{GAM}$	9.8700	152.9656	0.2172	0.0048	0.4779
	$\hat{T}_{LP}$	10.0203	156.9519	0.2205	0.0049	0.4852
n=8000	$\hat{T}_{NN}$	3.7104	20.7253	0.060	0.0011	0.1102
	$\hat{T}_{MARS}$	4.3909	29.5419	0.0779	0.0014	0.1382
	$\hat{T}_{GAM}$	10.4348	30.3327	0.4512	0.0041	0.4128
	$\hat{T}_{LP}$	13.1195	40.3936	0.2327	0.0080	0.8003

4.3 Unconditional Properties for Artificial Data

4.3.1 Unconditional Properties for the Two Dimensional Linear Model

The Generalized Additive Estimator and Local Polynomial Regression Estimator over Estimates the Finite Population Total under two dimensional model. This is because of their large Bias values of 9.8700 and 10.0203 respectively. The Finite Population Estimator $\hat{T}_{NN}$ has lower Biases, Mean Square Error, Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors which is followed closed by the Estimator $\hat{T}_{MARS}$. Therefore, the Estimator of Finite Population Total based on Feedforward Backpropagation Neural Network emerges the best and favorable in Estimating the Finite Population Total.

It is recorded and observed that, as the Sample size increases, all the Estimators recorded a significant improvement in their performance in Estimating the Finite Population Total. Notably is the Local Polynomial Regression Estimator with a significance reduction in Bias and Mean Square Errors. The Neural Network Estimator still outperforms al other Estimators with significant reduction in Biases, Mean Square Error Relative Root Mean Square errors, Mean Absolute Errors and Mean Absolute Percentage Errors as Sample sizes increases.

Table 4.2: Unconditional Bias, Mean Square Error, Relative root mean Square Error, Mean Absolute Error and Mean Absoute Percentage Error for Two Dimensional Quadratic Model

		Bias	MSE	RMSE	MAE	MAPE
n=4000	$\hat{T}_{NN}$	3.3743	20.5077	0.0596	0.0011	0.1052
	$\hat{T}_{MARS}$	6.8289	76.5408	0.1206	0.0021	0.2130
	$\hat{T}_{GAM}$	20.0105	643.8682	0.3534	0.0062	0.6240
	$\hat{T}_{LP}$	18.1960	536.6546	0.3213	0.0057	0.5675
n=8000	$\hat{T}_{NN}$	1.9396	12.5319	0.0343	0.0006	0.0605
	$\hat{T}_{MARS}$	4.0274	25.4001	0.0711	0.0013	0.1256
	$\hat{T}_{GAM}$	12.4017	246.0122	0.2190	0.0039	0.3868
	$\hat{T}_{LP}$	11.3112	200.8425	0.1998	0.0035	0.3528

4.3.2 Unconditional Properties for the Two Dimensional Quadratic Model

Table 4.2 summarizes the results for the performance of the Estimators for a Two Dimensional Quadratic model. Compared to Linear case, the performance of all the Estimators has marginally decreased as indicated by marginal increase Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors across all the Estimators of Finite Population Total.

It is also observed that, the Generalized Additive Estimator and Local Polynomial Regression records poor performance in terms of Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors in Estimating the Finite Population Total under two dimensional Quadratic Model. This is because of their large Bias values of Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors. The finite population estimator $\hat{T}_{NN}$ has lower Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors which is followed closed with the Estimator $\hat{T}_{MARS}$.

Therefore, the Estimator of Finite Population Total based on FeedForward Backpropagation Neural Network emerges the best and favorable in Estimating the finite population total in the two dimensional Quadratic case.

Even with the Sample increases, all the Estimators record a significant improvement in their performance in estimating the Finite Population Total. To note here is the Local Polynomial Regression Estimator with a significance reduction in Bias and Mean Square Errors. The Neural Network Estimator still outperforms other Estimators as significant reduction in Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors as Sample sizes increases.

4.3.3 Unconditional Properties for the Three Dimensional Mixed Model

Table 4.3: Unconditional Bias, Mean Square Error, Relative root mean Square Error, Mean Absolute Error and Mean Absoute Percentage Error for Three Dimensional mixed Model

		Bias	MSE	RMSE	MAE	MAPE
n=4000	$\hat{T}_{NN}$	3.5196	18.9278	0.0583	0.0010	0.0965
	$\hat{T}_{MARS}$	5.7422	52.2492	0.0951	0.0016	0.1574
	$\hat{T}_{GAM}$	14.7975	353.5178	0.2450	0.0041	0.4056
	$\hat{T}_{LP}$	16.8233	437.2852	0.2785	0.0046	0.4612
n=8000	$\hat{T}_{NN}$	1.8147	5.4731	0.0300	0.0005	0.0497
	$\hat{T}_{MARS}$	3.3823	18.3560	0.0560	0.0009	0.0927
	$\hat{T}_{GAM}$	8.8086	122.7989	0.1458	0.0024	0.2415
	$\hat{T}_{LP}$	9.9900	151.4552	0.1654	0.0027	0.2738

Table 4.3 summarizes the results for performance of the Estimators for a three dimensional mixed model. Compared to two dimensional case, the performance of all the estimators has marginally decreased as indicated by marginal increase Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors across all the Estimators of Finite Population Total.

It is also observed that, the Generalized Additive Estimator and Local Polynomial Regression still recorded poor performance in terms of Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors in Estimating the Finite Population Total under three dimensional mixed model. This is because of their large values of Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors.

In the other case, the Finite Population Estimator $\hat{T}_{NN}$ has lower Biases, Mean Square Error Relative Root Mean Square Errors, Mean Absolute Errors and Mean Absolute Percentage Errors which is followed closed with the Estimator $\hat{T}_{MARS}$. Therefore, the Estimator of Finite Population Total based on Feedforward Neural Network emerges the best and favorable in Estimating the Finite Population Total in the three dimensional mixed model case.

Even with the Sample increases, all the Estimators record a significant improvement in their performance in Estimating the Finite Population Total. To note here is the Local polynomial Regression Estimator with a significance reduction in Bias and Mean Square Errors. The Neural Network Estimator still remains the Estimator of choice compared to other Estimators as Sample sizes increases.

4.4 Conditional Properties for Artificial Data

The 1000 Simple Random Samples were sorted using the Sample Means of $\bar{X}_s$ values criterion. The Samples were then grouped into sets of twenty Samples such that the first set is made of Samples with the lowest Sample Means of $\bar{X}_s$ values, the second set consists of Samples with Means of $\bar{X}_s$ that are larger than the Sample Means of the first set and so on until the last set that consists of Samples with the largest Sample Means of $\bar{X}_s$ values. In each of the group, the Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error were computed.

The results of group the Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error for the Finite Population Total Estimators $\hat{T}_{NN}$, $\hat{T}_{MARS}$, $\hat{T}_{GAM}$ and $\hat{T}_{LP}$ are plotted against group average values $\bar{\bar{X}}$ denoted as Xbar in the fifty groups of Mean of $\bar{X}_s$.

Figures 4.2-4.5 summarizes the findings of the Conditional results for the Estimators under Two dimensional Linear model, Two Dimensional Quadratic model and Three Dimensional Mixed Model respectively. From the results, It was observed that both the Estimators overestimates the Finite Population Total. However, Generalized Additive and Local Polynomial Estimators performs poorly compared to Neural Network and MARS Estimators of Finite Population Total. The Neural Network Estimator emerges to perform better across all the models hence becoming the most preferred Estimator in high dimensional dataset.

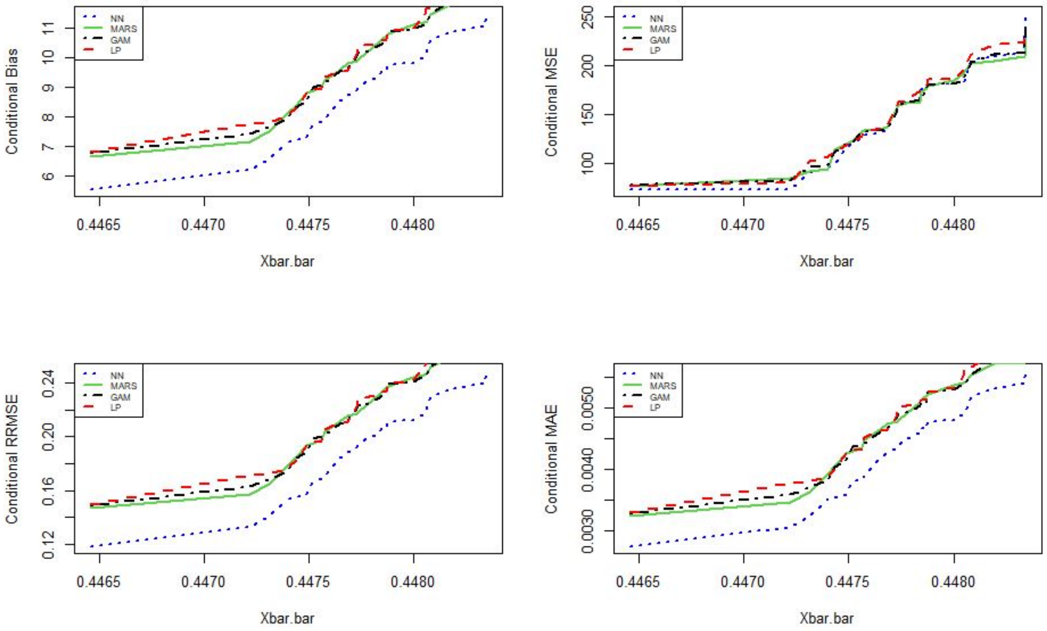

Figure 4.2: Conditional Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error based on a Two dimensional Linear model

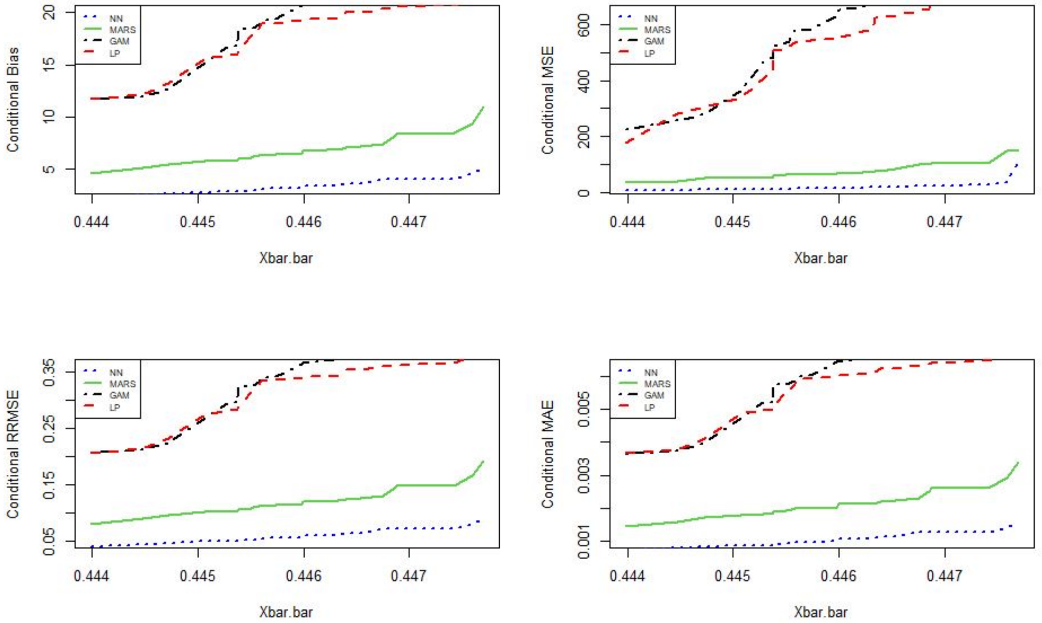

Figure 4.3: Conditional Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error based on a Two dimensional Quadratic model with sample size 4000

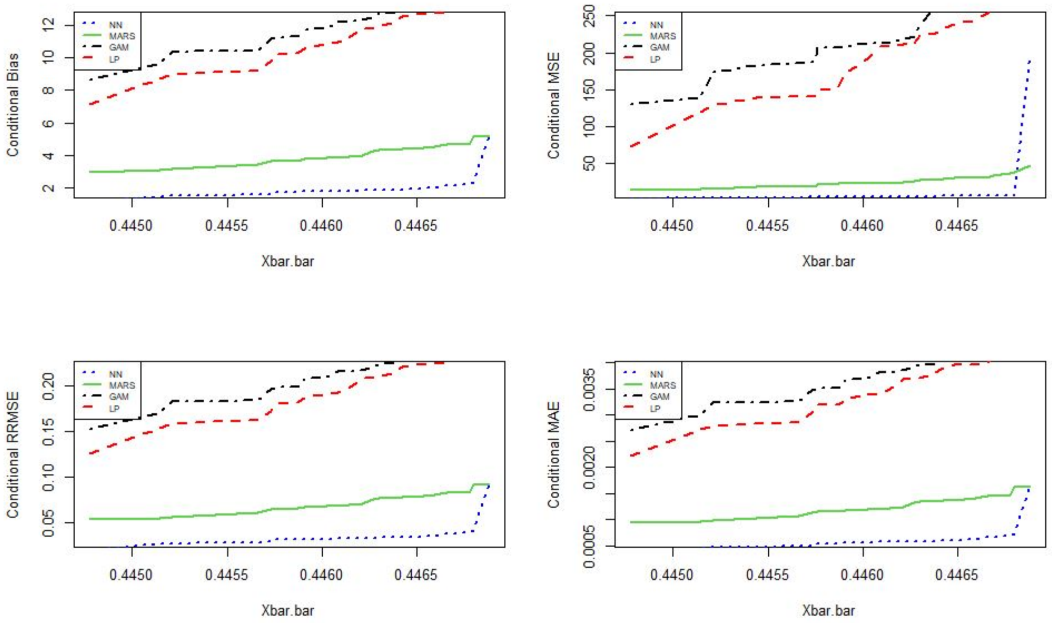

Figure 4.4: Conditional Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error based on a Two dimensional Quadratic model with sample size 8000

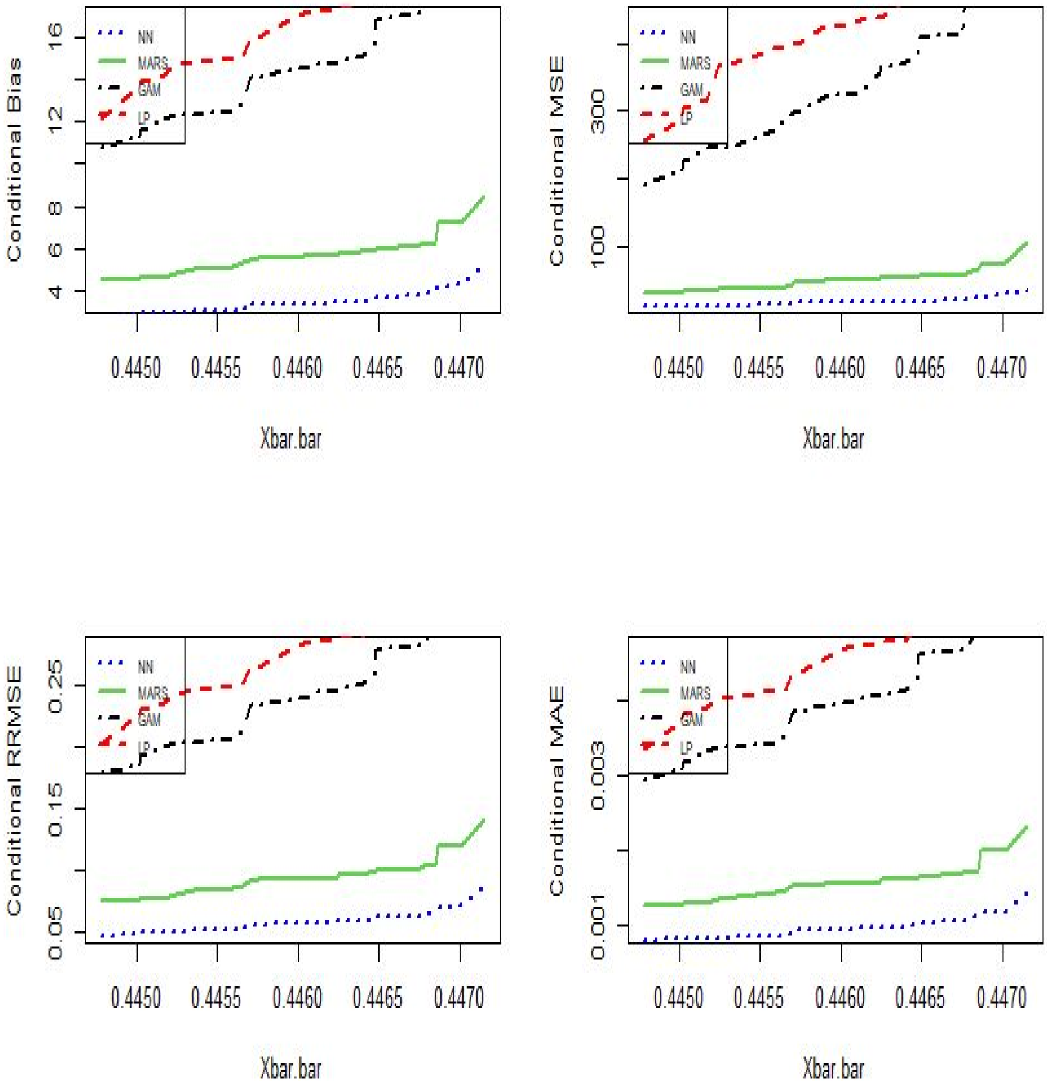

Figure 4.5: Conditional Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error based on a Three dimensional mixed model with a sample size of 4000

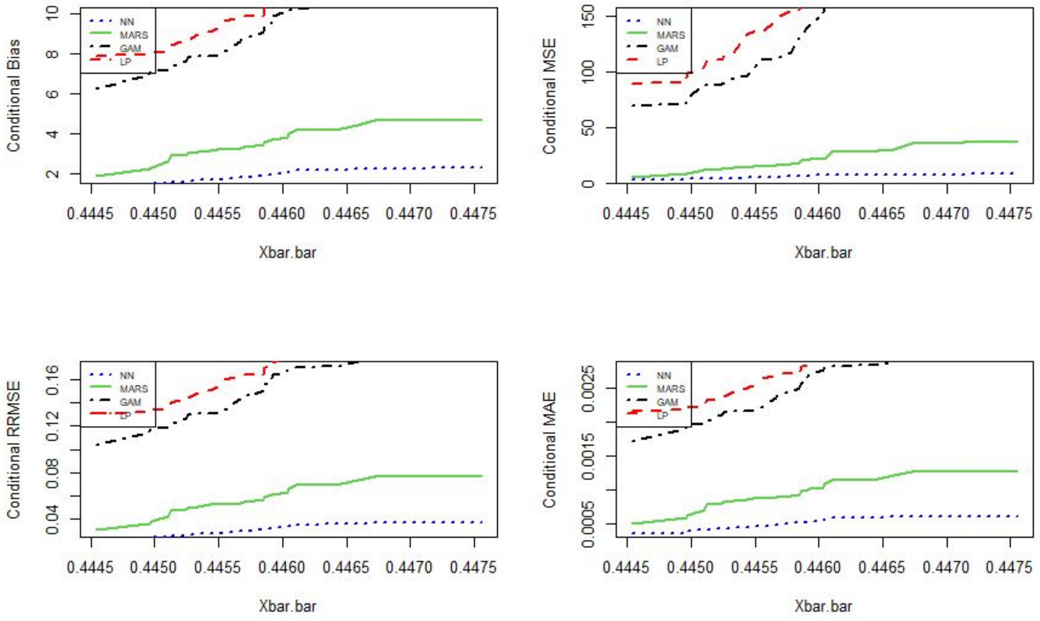

Figure 4.6: Conditional Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error based on a Three dimensional mixed model with a sample size of 8000

4.5 Application of the Developed Estimator to the Data from United Nations Development Programme

To illustrate our Estimation approach, the following data was utilized. A Population of size 189 was obtained from the United Nations Development Programme 2020 report. The UN studied the development in 189 countries. It grouped development in the countries as either very high human development, high human development, medium human development or low human development. Kenya was classified in countries that falls under medium development and ranked number 143 out of the 189 countries studied.

The UN study used attributes such as Human Development Index(HDI), Life expectancy at Birth,Expected years of schooling, Mean years of schooling, Gross national income (GNI) per capita and GNI per capita rank minus HDI to rank human development index in the 189 countries. In this study, a relationship between Human Development Index(HDI) which is considered as the survey variable and the auxiliary variables ; Life expectancy at Birth, Expected years of schooling, Mean years of schooling and Gross National Income (GNI) per capita was considered.

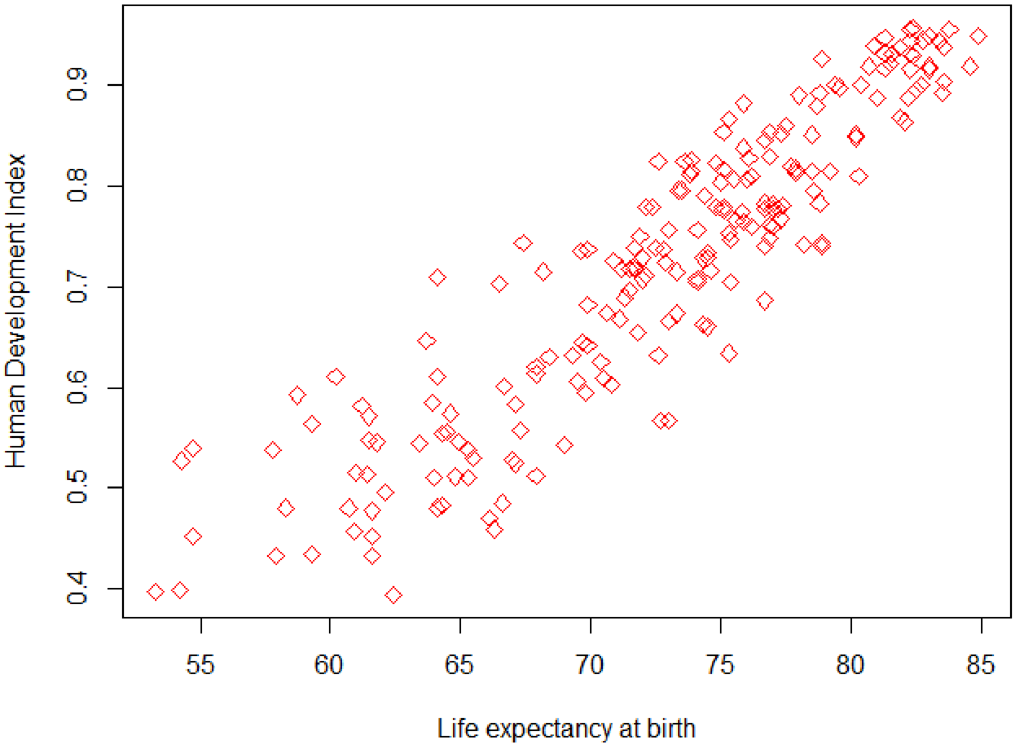

Figure 4.7: Relationship Between Human Development Index(HDI) and Life Expetancy at Birth(LEB)

From the Scatter Plot in Figure 4.7, a Quadratic relationship between HDI and LEB was observed. This indicates a strong positive relationship between LEB and HDI.

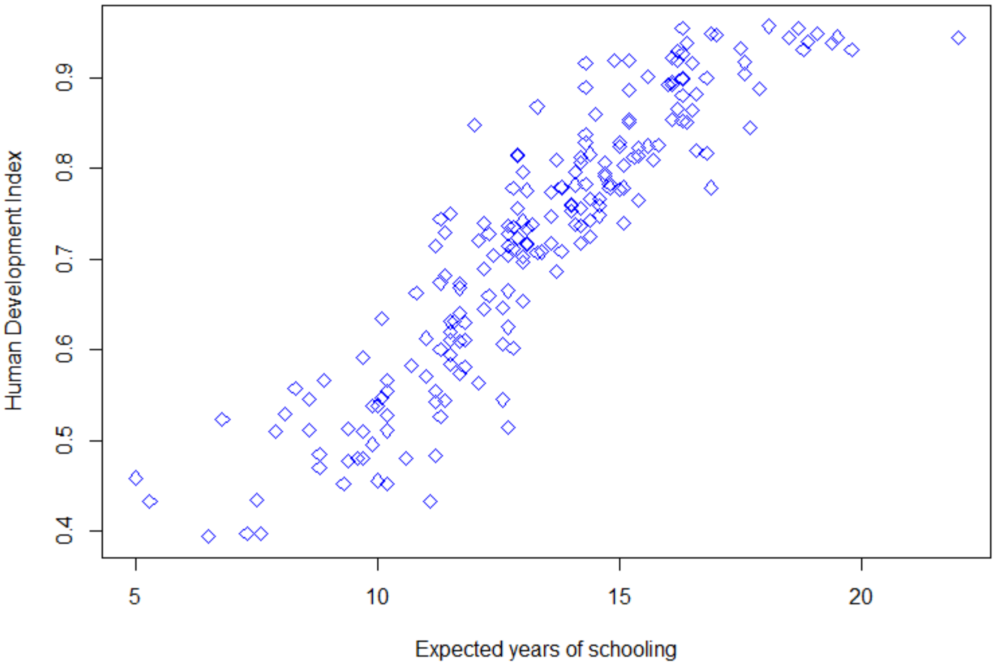

Figure 4.8: Relationship Between Human Development Index(HDI) and Expected years of schooling (EYS)

From the scatter plot in Figure 4.8, a Quadratic relationship between HDI and EYS was observed. This indicates a strong positive relationship between EYS and HDI. The graphs also shows the presence of outliers in the data which might affect the efficiency of Parametric models when assumed in the analysis of this data.

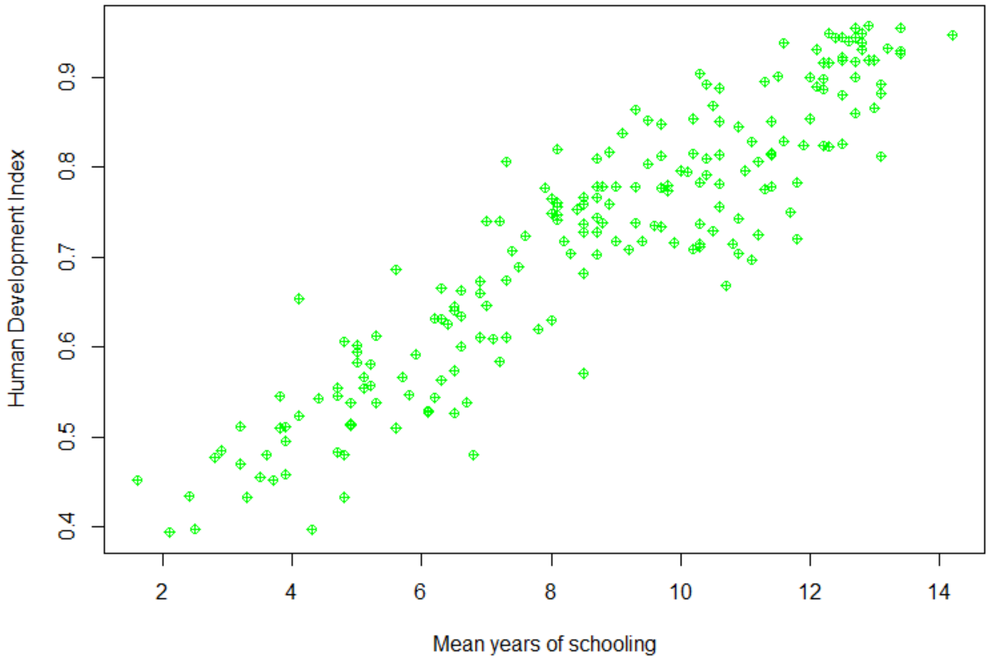

Figure 4.9: Relationship Between Human Development Index(HDI) and Mean years of schooling (MYS)

From the Scatter Plot in Figure 4.9, a Linear Relationship between HDI and MYS was observed. This indicates a strong positive relationship between MYS and HDI.

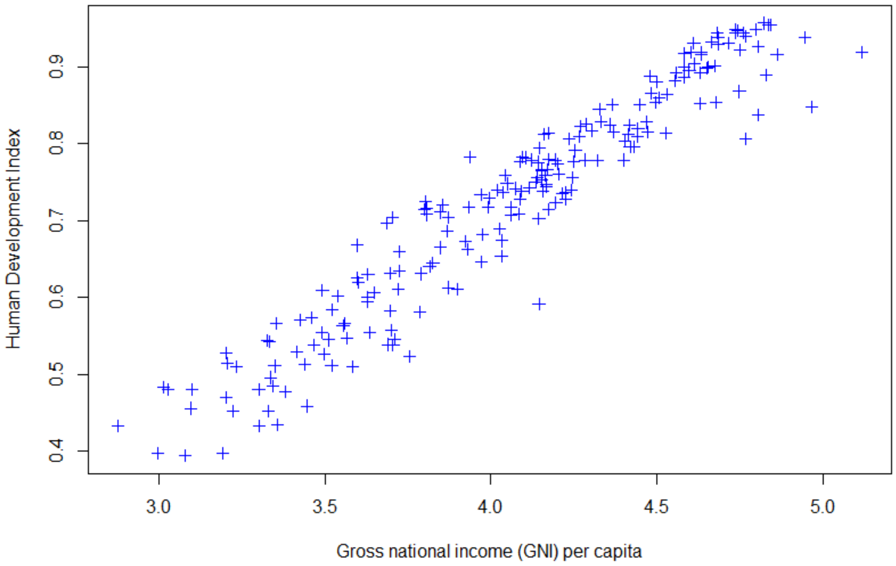

Figure 4.10: Relationship Between Human Development Index(HDI) and Gross National Income (GNI) per capita

From the scatter plot in Figure 4.10, a Linear relationship between HDI and GNI was observed. This indicates a strong positive relationship between GNI and HDI. The graphs also shows the presence of outliers in the data which might affect the efficiency of Parametric Models when assumed in the analysis of this data.

To study the performance of the developed Estimator in relation to other Estimators considered for comparison in this study, two Samples of of size 50 and 100 were taken from Population (United Nations Development Programme 2020) applying Simple Random Sampling without replacement design. For each sample taken, Bias, Mean Suare Error, Relative Mean Error, Mean Absolute Error and Mean Absolute Percentage Error were computed.

Table 4.4: Unconditional Bias, Mean Square Error, Relative root mean Square Error, Mean Absolute Error and Mean Absolute Percentage Error for Real Data Set

		Bias	MSE	RMSE	MAE	MAPE
n=50	$\hat{T}_{NN}$	0.0289	0.0013	0.0023	0.0001	0.0132
	$\hat{T}_{MARS}$	0.0541	0.0046	0.0043	0.0003	0.0346
	$\hat{T}_{GAM}$	0.0580	0.0052	0.0046	0.0004	0.0371
	$\hat{T}_{LP}$	0.0331	0.0017	0.0026	0.0002	0.0211
n=100	$\hat{T}_{NN}$	0.0145	0.0003	0.0012	0.0001	0.0103
	$\hat{T}_{MARS}$	0.0279	0.0012	0.0022	0.0002	0.0178
	$\hat{T}_{GAM}$	0.0319	0.0016	0.0025	0.0002	0.0204
	$\hat{T}_{LP}$	0.0184	0.0005	0.0015	0.0001	0.0118

Table 4.4 shows the Estimated Bias, Mean Square Error, Relative Mean Error, Mean Absolute Error and Mean Absolute Percentage Error for each Estimator considered. From these results, it can be observed that the overall performance of the T_{NN} estimate is superior to the usual one since it has minimum Bias, MSE, Relative Mean Error and Mean Absolute Error. Even as the sample increases, all the Estimators recorded a significant improvement in their performance in Estimating the Finite Population Total. The developed Estimator, T_{NN} still outperforms other Estimators as significant reduction in Biases, Mean Square Error, Relative Mean Error and Mean Absolute Errors was noticed as Sample sizes was increased.

The conditional performance of the Estimator was done and compared with the performance of other existing Estimators of Finite Population Total . To do this, 500 random samples, all of sizes 50 and 100, were selected and the Mean of the Auxiliary values x_i was computed for each Sample to obtain 200 values of $\bar{X}$. These Sample Means were then sorted in ascending order and further grouped into clusters of size 20 such that a total of 25 groups was realized.

Further, group means of the Means of Auxiliary variables was calculated to get $\bar{\bar{X}}$. Empirical means and Biases were then computed for all the Estimators T_{NN}, T_{LP}, T_{MARS} and T_{GAM}. The Conditional Biases were plotted against $\bar{\bar{X}}$ to provide a good understanding of the pattern generated. Figure 4.11 and 4.12 summarizes the behaviors of the Conditional Biases, Relative Absolute Biases and Mean Squared Error realized by all the Estimators.

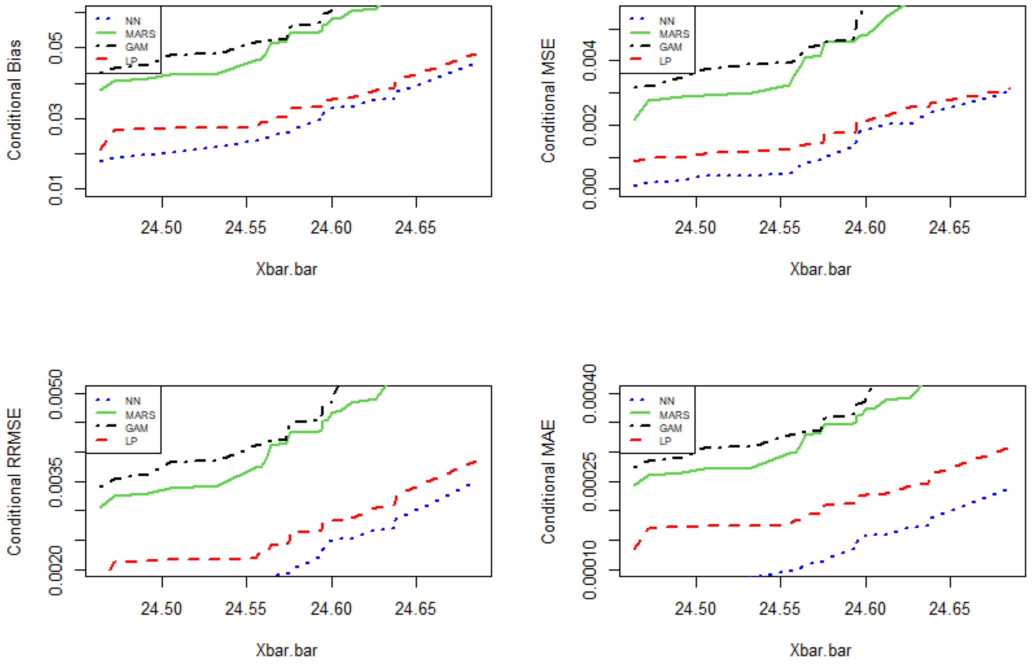

Figure 4.11: Conditional Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error based on real data with a sample size of 100

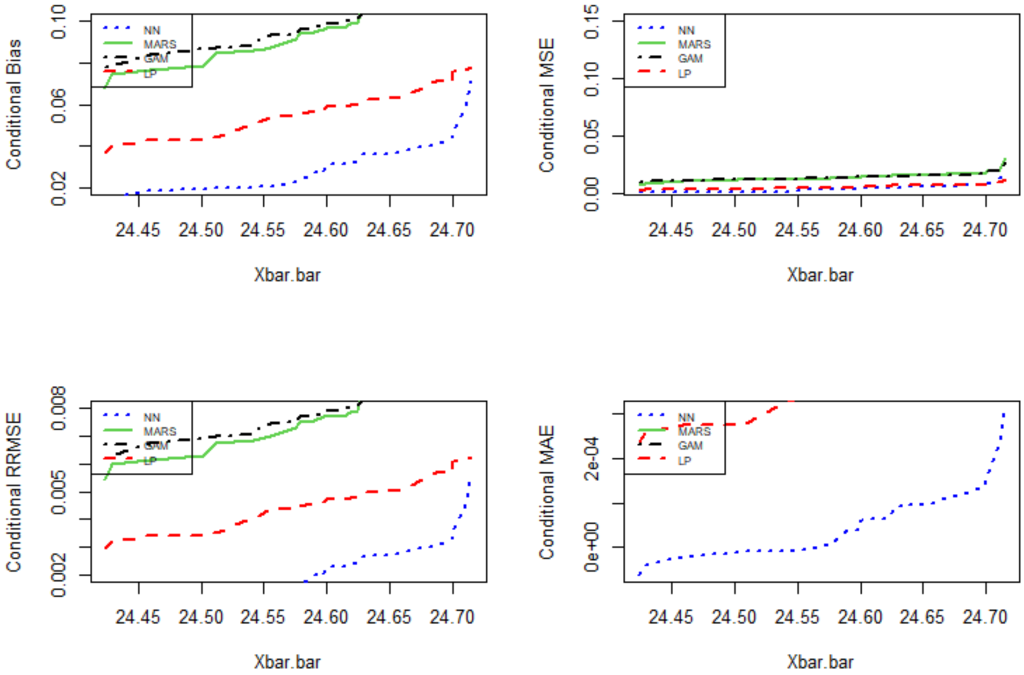

Figure 4.12: Conditional Bias, Mean Square Error, Relative Root Mean Square Error and Mean Absolute Error based on real data with a sample size of 50

In most cases there are significant differences among the Bias characteristics of the various Estimators. A detailed examination of the plots reveals that T_{NN} has a lower levels of Bias followed by T_{LP} as indicated by the proximity of plotted curves to the horizontal (no Bias) line at 0:0 on the vertical axis. Interestingly, despite the rather entangled nature of some of the plots, Estimator T_{NN} emerges clearly as the least Biased for nearly every group Means of the Means of Auxiliary variables.

From the plots, It can be observed that both the estimators underestimate the Finite Population Total. Plots of Conditional MSE versus group Means of the Means of Auxiliary variables similarly reveal coincident behavior for the Estimators. T_{NN} and T_{LP} produce generally the lowest MSE values. In particular, T_{NN} yields the lowest MSE in most cases among all other Estimators. T_{NN} is consistently better than all other Estimators for both Bias and MSE. All of these Estimators are Asymptotically Unbiased and they all exhibit MSE consistency in that the MSE values tend toward zero as sample size increases. From the plots it can be seen that T_{NN} and T_{LP} performed equally better than all other estimators of the true Population Total.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

In this final Charpter, a brief summary of key results from this thesis are presented. In doing so, a summary that expounds on what the thesis aluded to at the abstract stage has been provided followed by conclusions on each specific objective and ensuing recommendations guided by the theoretical an empirical analysis in this thesis. From all these, it is finally concluded that Estimator of Finite Population Total based on the Feedforward Backpropagation Neural Network has proved to yield results with great precision and therefore it is recommended for estimating Finite Population Total in High Dimensional Datasets in different sectors of the economy since it yields very good results.

5.2 Summary

The main goal in Survey sampling is to use the sample statistics to make conclusions about the overall Finite Population Quantities. Nonparametric regression has developed into a increasingly growing field of statistics providing a versatile and data analytical way of estimating Regression Function without specifying a Parametric Model correctly. The estimates from these approach are often more reliable and versatile than design based presumptions or Parametric Regression Models.

The Nonparametric Regression estimators suggested in literature contribute to the Trade-Off of Bias-Variance along the boundary points and hence becomes infeasible in high dimensions. It is because of these, the foundation of this Thesis is developed with the aim of addressing this weaknesses by applying Neural Network method to the estimation of the Finite Population Total in circumstances where one is dealing with high dimensional datasets.

Therefore In realizing the set out objectives, this thesis has demonstrated that it is possible obtain robust Nonparametric estimator of Finite Population Total in high dimensional datasets using the Feedforward Backpropagation Neural Network. The performance of this Nonparametric Feedforward Backpropagation Neural Network was found to be better when compared with GAM, MARS and LP estimators which was in line with analysis done by Montanari and Ranalli (2005) but in the univariate case. This is therefore a milestone when used in area of Survey sampling and specifically Nonparametric estimation when dealing with multivariate datasets which often the case in real life problems .

Globally data is needed to make decisions thus making Census in important way of data collection that plays a crucial role during resource allocation and planning. However, Census are carried out only after every ten years which limits there impact within the intermediate years. Thus, other methods are required for planning in the intervening years. Population Estimates use the Census as a baseline, for instance adding Births and subtracting Deaths and make allowances for Migration. They can be used for National and Local Planning. Population Estimates are produced annually. Additionally, National Government use Population Estimates as the basis for capitation-based funding of County Governments and Primary Care, Education, Health Sector Trusts, hence under-estimation can therefore have effects on Local services, and Over-estimation can lead to unfair resource distribution. Therefore having an Robust Estimator Population Total/ Estimates will ensure equitable resource allocation.The outcome of this thesis will play an important role in providing a reliable Estimator of Finite Population Total that can be used different fields of the economy.

Additionally, the study contributes towards development of Mathematical and Statistical knowledge in Survey Sampling. The developed Estimation Procedure is useful to policy makers since National Development is dependent on the Sampling Strategy employed. In addition, Business and Industrial sectors stand to benefit from this study by using the developed Estimation Procedure for prediction and thereby improving the efficiency of their internal operations.

5.3 Conclusions

The main goal of this thesis was to estimate Finite Population Total in high dimensional datasets based on a Robust Nonparametric Feedforward Backpropagation Neural Network technique. In this thesis, a Robust Nonparametric Feedforward Backpropagation Neural Network estimator of Finite Population Total was developed by employing a FeedForward Backpropagation Neural Network technique in Nonparametric Regression. Asymptotic properties such as the Consistency and Mean Squared Error for the developed estimator have been derived.

The first objective of this thesis was to develop a Robust Nonparametric Estimator of Finite Population Total based on Feedforward Backpropagation Neural Network within the model based approach. This study developed an estimator of Finite Population Total based on Feedforward Backpropagation Neural Network as given in equation 3.36 in section 3.5.

The second objective of this was to derive the asymptotic properties of the developed Robust Nonparametric Estimator of Finite Population Total based on Feedforward Backpropagation Neural Network. The asymptotic properties of the developed estimator were derived as provided in section 3.6, 3.7.2 and 3.7.3 of this thesis. By investigating properties of the developed estimator, it was concluded that it has Asymptotic Normal Distribution, as well as being Asymptotically Unbiased and Asymptotically Consistent Estimator of the Population Total.

The third and last objective of thesis thesis was to study the coverage properties of the developed Robust Nonparametric Estimator of Finite Population Total based on Feedforward Backpropagation Neural Network by comparing its performance to that of identified Nonparametric Finite Population Estimators using data from UNDP report and artificial data Simulated from certain models.

When applied to simulated data and dataset obtained from the United Nations Development Programme 2020 report, the findings indicate that the proposed estimator has the lowest bias and root mean square error values compared to other existing estimators such as multivariate adaptive regression splines (MARS), generalized additive model (GAM), and local polynomial (LP) which can handle high-dimensional data. As evidenced from the analysis of the Biases and Mean Square Errors presented in Tables 4.1, 4.2 and 4.3, it was possible to significantly reduce the Bias and increase precision. The Biases indicate that the proposed estimator is superior to the other estimators in all the models used in high dimensional cases.

The graphs of the Conditional Biases, Relative Biases and Mean Square Error in Figures 4.2, 4.3 and 4.4 also indicate that the proposed estimator dominates the other estimators. The graphs show that while the other estimators have larger Conditional Biases, the proposed estimator is almost Conditionally Unbiased. This good performance of the developed estimator was also evident with the Conditional Mean Square Error graphs.

Additionally, the following observations were made using both theoretical and empirical results;

(i) The Neural Network estimator estimates the finite population total better than all other robust estimators in high dimensional case.

(ii) The Performance of local polynomial estimator in the estimation of finite population becomes poor as the dimension of the data increases.

(iii) For all the estimators, as the sample sizes increases, Biases, Mean Square Error Relative root mean Square errors, Mean Absolute Errors and Mean Absolute Percentage Errors decreases for the four models considered.

(iv) For all the estimators, as the dimension increases, Biases, Mean Square Error Relative root mean Square errors, Mean Absolute Errors and Mean Absolute Percentage Errors decreases for all the four models considered.

The main conclusion in this thesis is that the Estimator of Finite Population Total based on the Feedforward Backpropagation Neural Network has proved to yield results with great precision and therefore it is recommended for estimating Finite Population Total in High Dimensional Datasets in different sectors of the economy since it yields very good results.

5.4 Recommendations

In this study, the assumption made was that the activation function was a Sigmoid function. Future research could examine usage of other activation functions which are not Sigmoid in nature such as Rectified linear unit (RecRELU) and the performance of the resulting estimator compared to see if it produces superior Finite Population Total estimation.

The estimator in this study has been considered in the case of simple random sampling without replacement (SRSWoR). Therefore, future research could examine extending this to other complex sampling techniques that rely on SRSWoR.

REFERENCES

Asnaashari, A., McBean, E. A., Gharabaghi, B., and Tutt, D. (2013). Forecasting watermain failure using artificial neural network modelling. *Canadian Water Resources Journal*, 38(1):24–33.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

Bickel, P. J. and Li, B. (2007). Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pages 177–186.

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

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, pages 1026–1053.

Brewer, K. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5(3):93–105.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Chambers, R., Dorfman, A. H., and Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, 79(3):577–582.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3):597–604.

Cochran, W. G. (1977). Sampling techniques-3.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Despagne, F. and Massart, D.-L. (1998). Variable selection for neural networks in multivariate calibration. *Chemometrics and intelligent laboratory systems*, 40(2):145–163.

Di Ciaccio, A. and Montanari, G. (2001). A nonparametric regression estimator of a finite population mean. *Book of Short Papers*, pages 173–176.

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

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

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Feng, J. and Simon, N. (2017). Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*.

Franke, J. and Diagne, M. (2006). Estimating market risk with neural networks. *Statistics & Decisions*, 24(2):233–253.

Franke, J. and Neumann, M. H. (2000). Bootstrapping neural networks. *Neural computation*, 12(8):1929–1949.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.

Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192.

Gasser, T. and Engel, J. (1990). The choice of weights in kernel regression estimation. *Biometrika*, 77(2):377–381.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing techniques for curve estimation*, pages 23–68. Springer.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, pages 401–414.

Githinji.S (2010). *Model Based Non-parametric Regression Estimation of Finite population Total under Two-Stage Cluster Sampling.* PhD thesis, Kenyatta University.

Godambe, V. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 269–278.

Grenander, U. and Ulf, G. (1981). Abstract inference. Technical report.

Ham, J., Lee, D. D., Mika, S., and Schölkopf, B. (2004). A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, page 47. ACM.

Hansen, M. H. et al. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2(2):180–190.

Härdle, W. and Linton, O. (1994). Applied nonparametric methods. *Handbook of econometrics*, 4:2295–2339.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., et al. (2004). *Nonparametric and semiparametric models*, volume 1. Springer.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press.

Herbert, I. O., Orwa, G. O., and Otieno, R. O. (2017). Optimal nonparametric regression estimation of finite population total using nadaraya watson incorporating jackknifing. *International Journal of Theoretical and Applied Mathematics.*

Holt, D. and Smith, T. (1979). Post stratification. *Journal of the Royal Statistical Society. Series A (General)*, pages 33–46.

Hwang, J. G. and Ding, A. A. (1997). Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 92(438):748–757.

Levina, E. and Bickel, P. J. (2004). Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784.

Liang, F. and Kuk, Y. (2004). A finite population estimation study with bayesian neural networks. *Survey Methodology*, 30(2):219–234.

Maier, H. R. and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental modelling & software*, 15(1):101–124.

Manzoor, M. A., Akbar, A., and Ullah, M. A. (2013). Performance of nonparametric regression estimation with diverse covariates. *Pakistan Journal of Social Sciences (PJSS)*, 33(1).

Montanari, G. E. and Ranalli, G. M. (2005). Nonparametric methods for sample surveys of environmental populations.

Montanari, G. E. and Ranalli, M. G. (2003). On calibration methods for design based finite population inferences. *Bulletin of the International Statistical Institute, 54 th session*, 60.

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

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, pages 134–153.

Opsomer, J. D., Breidt, F. J., Moisen, G. G., and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*, 102(478):400–409.

Otieno, R., Mwita, P., and Kihara, P. (2007). Nonparametric model assisted model calibrated estimation in two stage survey sampling. *East African Journal of Statistics*, 1(3):261–281.

Otieno, R. O. and Mwalili, T. M. (2000). Nonparametric regression method for estimating the error variance in unistage sampling.

Priestley, M. and Chao, M. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 385–392.

Rady, E.-H. A. and Ziedan, D. (2014). Estimation of population total using nonparametric regression models. *Advances and Applications in Statistics*, 39(1):37.

Robinson, P. and Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 240–248.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

Särndal, C. E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67(3):639–650.

Serfling, R. J. (2000). Approximation theorems of mathematical statistics. 1980.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Shahin, M. A., Jaksa, M. B., Maier, H. R., et al. (2002). Artificial neural network based settlement prediction formula for shallow foundations on granular soils. *Australian Geomechanics: Journal and News of the Australian Geomechanics Society*, 37(4):45.

Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2019). Asymptotic properties of neural network sieve estimators. *arXiv preprint arXiv:1906.00875*.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Soh, Y., Hae, Y., Mehmood, A., Ashraf, R. H., Kim, I., et al. (2013). Performance evaluation of various functions for kernel density estimation. *Open J Appl Sci*, 3(1):58–64.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053.

Takezawa, K. (2005). *Introduction to nonparametric regression.* John Wiley & Sons.

Tsybakov, A. B. and Tsybakov, A. B. (2009). Nonparametric estimators. *Introduction to Nonparametric Estimation*, pages 1–76.

Wang, S. and Dorfman, A. H. (1996). A new estimator for the finite population distribution function. *Biometrika*, 83(3):639–652.

White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural networks*, 3(5):535–549.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.

Zheng, H. and Little, R. J. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19(2):99.

Zucchini, W., Berzel, A., and Nenadic, O. (2003). Applied smoothing techniques. *Part I: Kernel Density Estimation*, 15.