Open Journal of Statistics, 2022, 12, 581-596
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Estimation of Finite Population Totals in High Dimensional Spaces

Festus A. Were, George O. Orwa, Romanus O. Otieno

Department of Statistics and Actuarial Sciences, JKUAT, Nairobi, Kenya
Email: weref87@gmail.com, gorwa@buc.ac.ke, rodhiambo@must.ac.ke

How to cite this paper: Were, F.A., Orwa, G.O. and Otieno, R.O. (2022) Estimation of Finite Population Totals in High Dimensional Spaces. *Open Journal of Statistics*, **12**, 581-596.
https://doi.org/10.4236/ojs.2022.125035

Received: August 12, 2022
Accepted: October 8, 2022
Published: October 11, 2022

Abstract

In this paper, the problem of Nonparametric Estimation of Finite Population Totals in high dimensional datasets is considered. A robust estimator of the Finite Population Total based on Feedforward Backpropagation Neural Network is derived with the aid of a Super-Population Model. This current study is motivated by the fact that Local Polynomials and Kernel methods have in preceding related studies, been shown to provide good estimators for Finite Population Totals but in low dimensions. Even in these situations however, bias at boundary points presents a big challenge when using these estimators in estimating Finite Population parameters. The challenge worsens as the dimension of regressors increase. This is because as the dimension of the Regressor Vectors grows, the Sparseness of the Regressors' values in the design space becomes unfeasible, resulting in a decrease in the fastest achievable rates of convergence of the Regression Function Estimators towards the target curve, rendering Kernel Methods and Local Polynomials ineffective to address these challenges. This study considers the technique of Artificial Neural Networks which yields robust estimators in high dimensions and reduces the estimation bias with marginal increase in variance. This is due to its Multi-Layer Structure, which can approximate a wide range of functions to any required level of precision. The estimator's properties are developed, and a comparison with existing estimators was conducted to evaluate the estimator's performance using real data sets acquired from the United Nations Development Programme 2020. 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 Neural Network estimator is highly recommended for survey sampling estimation of the finite population total.

Keywords

Neural Networks, Kernel Smoother, Local Polynomial, Nonparametric

1. Introduction

In Surveys, extrapolation reduces the accuracy of information since the sample is a subset of an entire population and therefore, does not contain information on units that are not represented in the selected sample. In such cases of unobserved units therefore, use of Auxiliary Information on the characteristic under study is usually effective in predicting unobserved units if the model is correctly specified. In general, when using Auxiliary Information, it is assumed that there is a finite population of N distinct and identifiable units; $U=\{1,2,\cdots,N\}$. Let each population unit have the variable of interest as Y. It is assumed that there is an auxiliary variable $X\in\mathbb{R}^d$, closely correlated with Y, which is known for the entire population (*i.e.* $X_1,X_2,\cdots,X_N$) that is known as $\forall\, Y_i$. Researchers are frequently faced with the task of estimating a population function, (*i.e.* a function of Y's), such as the Population Total;

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

or the population distribution functions

$$F(y)=\frac{1}{N}\sum_{i=1}^{N}I_i\left(Y_i\le y\right) \tag{2}$$

In estimating the Population Totals T for instance, a sample S is usually chosen such that the pair $\left(x_{i,j},y_i\right), i=1,2,\cdots,n$ and $j=1,2,3,\cdots,d$ is obtained from the variable X and corresponding variable Y. It can then be employed in the design, estimation, or both stages. In the presence of such Auxiliary Variables, Super-Population Models at the estimation stage of inference may be used, [1] and [2]. However, regarding the underlying relationship between the Survey and Auxiliary Variables, all of these techniques refer to Simple Statistical Models (Linear Regression Models). In an Empirical Study, [3] show that misspecification of the model can lead to substantial mistakes in the Parametric Superpopulation. To solve this problem, Nonparametric Regression involving robust estimators in Finite Population Sampling has been proposed [4] [5] [6].

As a result, the reason for using a nonparametric approach in this research is that a regression curve created this way serves four key functions, as explained by [7]: It provides a versatile method of exploring the general relationship between two variables, enables one to make prediction of observations without any reference to fixed parametric model, is a tool for finding spurious observations by studying influence of isolated points and is a flexible method for interpolating between adjacent values of auxiliary variable.

Usually, a major problem that is 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, ([8]). It is also known that Kernel and Polynomial Regression Estimators provide good estimates for the population totals when $x\in\mathbb{R}^d$ and $d=1$, [5] [9].

Despite the fact that High Dimensional Auxiliary Information can be accounted for in the above estimators, the problem of Regressor Sparseness in the design space renders Kernel Methods and Local Polynomials unworkable because performance decreases quickly as the dimension increases, [9] [10] [11]. This problem is known as "**curse of dimensionality**" which is a result of the sparsity of data in high-dimensional environments, which leads to a drop in the highest feasible rates of convergence of regression function estimators towards their target curve as the dimension of the Regressor Vector grows. A review on the concept of curse of dimensionality is provided in [12].

Given the problem called "curse of dimensionality", one has to use different Nonparametric Estimators to retain a large degree of flexibility. An attempt to navigate through this curse while handling Multiple Auxiliary Information is to consider and use recursive covering in model based perspectives [13] and Generalized Additive Modeling in Model-Assisted Framework [14]. These estimation methods come at a cost of reduced flexibility with the associated risk of increased bias [10] [11] [12] [15].

Consequently in this paper, robustness of the proposed Nonparametric Estimator for the Finite Population Total is based on Feedforward Backpropagation Neural Network Approach to address the shortcomings of previously studied estimation methods is developed. Although Kernel and Local Approximators may also have the same property as Artificial Neural Networks (ANNs), they often require a high number of components to attain equivalent approximation accuracy [16]. The high number of components presents a challenge to feasibility in usage of the methods. ANNs are thus considered to be a parsimonious approach to this Parametric Functional Analysis.

2. Estimation of Finite Population Totals Using Artificial Neural Networks

Let Y be the Survey Variable associated with an Auxiliary Variable X assumed to follow a Superpopulation Model under a Model-Based Approach. A commonly used working model for the Finite Population is

$$y_i = m(x_i) + \varepsilon_i \tag{3}$$

with $x_{ij} \in \mathbb{R}^d$, $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N$ i.i.d with mean zero and $x_{ij}, i = 1, 2, \cdots, N; j = 1, 2, \cdots, d$ are the auxiliary information.

Also, let

$$T = \sum_{i \in s} y_i + \sum_{i \in r} y_i \tag{4}$$

be the finite population total where s is the sampled units and r are the non-sampled units. Assume that y_i is given according to Equation (3) with $x_i \in \mathbb{R}^d$, $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N$ *i.i.d.* Consider estimating $m(x)$ based on a Feedforward Backpropagation Neural Network. As a basic building block, consider the Neurons as a Nonlinear Transformation of a Linear Combination of the input $x = (x_1, \cdots, x_d)'$.

Feedforward networks with multiple layers of hidden units are more complex networks that enable information feedback to be specified. Its study will only deal with the presented structure 5, which is widely used for a range of applications and has the appealing characteristic of being implemented in statistical software, but the results herein are straightforward to extrapolate.

In this simplest case of one hidden layer with $H \geq 1$ Neurons, the Network can be written to represent the Network Function as follows

$$f_H(x,\theta) = v_0 + \sum_{h=1}^{H} v_h \psi\left(w_{0h} + x^{\mathrm{T}} w_h\right),\ x \in \mathbb{R}^d \tag{5}$$

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

$$\theta = \left(w_{01}, \cdots, w_{0H}, w_1^{\mathrm{T}}, \cdots, w_H^{\mathrm{T}}, v_0, \cdots, v_h\right)^{\mathrm{T}} \in \mathbb{R}^{M(H)} \tag{6}$$

where $M(H) = (d+1)H + H + 1$ represents the vector of all parameters of weights of the network. $\psi : \mathbb{R} \mapsto \mathbb{R}$ is a given Activation Function. For regression problems, functions of the sigmoid shape. Therefore, 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 map into probability space. In particular, the Activation Function is a smooth counterpart of the Indicator Function if the input signals are "constrained" between zero and one. For instance, logistic function described as

$$\psi(u) = \frac{1}{1+\exp(-u)},\ -\infty < u < \infty \tag{7}$$

is a leading example of which it approaches one (zero) when its arguments go to infinity (negative infinity). Thus, the Logistic Activation Function produces partially on/off signals following the received input signals. This function $f_H(x;\theta)$ specifies a mapping from the input space $\mathbb{R}^d$ to the output space which for this study is one-dimensional. Such a class of all network output function
$O = \left\{ f_H(x;\theta), \theta \in \mathbb{R}^{M(H)}, H \geq 1 \right\}$ has several uniform approximation properties [17] [18] [19]. Important for the current study is that 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

$$\sup_{x \in C} \left| m(x) - f_H(x;\theta) \right| < \varepsilon$$

These imply that any Regression Function $m(x)$ may be approximated well enough using a large enough number of neurons and appropriate parameters θ.

Therefore, a nonparametric estimate for $m(x)$ is gotten if H is first chosen in a manner which serves as a tuning parameter and determines the smoothness of the estimate, then estimation of the parameter θ from the data by nonlinear least squares is done to yield

$$\hat{\theta}_n = \arg \min_{\theta \in \Re^{M(H)}} D_n(\theta) \tag{8}$$

with

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

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

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

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 $m(x)$. Therefore, the immediate consequence of these is that $f_H(x;\hat{\theta}_n)\to f_H(x;\theta)$ as $n\to\infty$.

The estimation error $\hat{\theta}_n-\theta$ can be divided into two asymptotically independent subcomponents: $\hat{\theta}_n-\theta=(\hat{\theta}_n-\hat{\theta}_n)+(\hat{\theta}_n-\theta)$, where the value

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

minimises the sample version of $D(\theta)$, [20]. Thus, by Universal Approximation Property of Neural Networks, $f_H(x;\theta)$ converges to the Regression Function $m(x)$ as $H\to\infty$. Therefore $f_H(x;\hat{\theta}_n)$ is a consistent Estimate of $m(x)$ if H increases with n as is herein imposed, and with an appropriate rate. From these results, the corresponding estimate of the finite population total is therefore, given as

$$\hat{T}_{NN}=\sum_{j\in s}y_j+\sum_{j\in r}\hat{m}_n(x_j)\tag{9}$$

which is the proposed estimator for the Finite Population Total, with

$$\hat{m}_n(x_j)=f_H(x;\hat{\theta}_n)$$

Regularity Notes on the Proposed Estimator

1) 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.

2) This estimator is identical to that proposed in [4], except that the NN is replaced by a Kernel-Based Regression.

3) This estimator can be used to estimate the population totals of a finite population so long as the assumption is that each of the unsampled elements has the same distribution as the sampled elements.

4) For fixed *H*, this work just fits a Nonlinear Regression Model to the data. However, it is known that this model can be misspecified and therefore one has 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 $m(x)$.

5) The parameter vector θ of [5] is not uniquely determined (identified) by the function $f_H(x,\theta)$. *i.e.* for different values of θ, the same function $f_H(x,\theta)$

is realised. If, for example the activation function is antisymmetric, $\psi(-x) = -\psi(x)$, then changing the enumeration of hidden units and multiplying all weights $w_{ih}, i = 1, 2, \cdots, d$, going into hidden units and simultaneously the weight v_h going out of the neuron by −1 do not change the function. To avoid this ambiguity and the related problems of estimation, this study considered only parameter vectors in a subset $\Theta_H \subset \mathbb{R}^{M(H)}$ chosen such that for each function in [5] with H neurons, there exists exactly one corresponding parameter Θ_H. For antisymmetric ψ one can choose for example $\Theta_H = \left\{\theta \in \mathbb{R}^{M(H)}; v_1 \geq v_2 \geq \cdots \geq v_H\right\}$, that is, the last h coordinates of θ are in decreasing order. For more details on the identification of parameters see [21].

Theoretically, Feedforward Neural Network which has one hidden layer suffices by the Universal Approximation Property. For practical purposes, networks with more than one hidden layer may provide a better approximation to $m(x)$ with fewer parameters, see [9] [17] [18] [22] [23].

3. Theoretical Properties of the Proposed Estimator

3.1. Assumptions

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

1) The errors ε_i are Identically Independently Distributed (IID) 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 $\alpha > 0$.

2) The Auxiliary Measurements $x_i \in \mathbb{R}^d$ are i.i.d. with an absolutely continuous distribution Fhaving a finite second moment.

$$\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_d} f\left(t_1,\cdots,t_d\right)\mathrm{d}t \tag{10}$$

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

$$pr\left(\|x_i\| > t\right) \leq b_0 \exp\left\{-b_1 t^{\beta}\right\} \text{ for all } t \geq 0 \tag{11}$$

and for some b_0, b_1 and $\beta > 0$.

3) $m(x)$ is a bounded function.

4) For each sequence of finite population indexed by *v*, conditioned on the value x_i, the super population model (3.1), where ε_i satisfies A1, then, the x_i is considered fixed with respect to the super population model ξ.

5) The survey variable has a bounded moment with *ξ-probability* 1. Moreover, it is noted that (A1), …, (A3) immediately imply for some $c_0, c_1 > 0$

$$Pr\left(|y_i| > t\right) \leq c_0 \exp\left\{-c_1 t^{\alpha}\right\}, \text{ for all } t \geq 0 \tag{12}$$

6) The sampling rate is bounded, that is

$$\limsup_{v\to\infty} \frac{n}{N} = \pi, \text{ where } \pi \in (0,1)$$

7) 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 [21].

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

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

c) $\left(w'_h, w_{0h}\right) = \pm\left(w'_{h'}, w_{0h'}\right)$, for $w \neq w'$.

8) The activation function ψ in 7 is asymmetric sigmoid function that is differentiable to any order. Additionally, it is assumed that the class of functions $\left\{\psi\left(b_t, b_0\right), b > 0\right\} \cup \left\{\psi \equiv 1\right\}$ is linearly independent. Such function can easily be represented using an indicator (threshold) function,

$$\psi(u) = \begin{cases} \psi(u) \to 0, \text{ as } u \to -\infty \\ \psi(u) \to 1, \text{ as } u \to +\infty \\ \psi(u) + \psi(-u) = 1 \end{cases} \tag{13}$$

The logistic activation function in 7 fulfills these requirements.

To prove for consistency of the proposed estimator, the rate which determines how the complexity of the networks and therefore the possible roughness of the function estimate $\hat{m}_n(x)$ increases with the sample size n has to satisfy some conditions. We follow [19] and restrict the number H of neurons and the overall size of the network weights v_h, w_{kh} simultaneously. For some sequences $H_n, \Delta_n \to \infty$, let

$$\Theta_n = \Theta\left(H_n, \Delta_n\right) = \left\{\theta \in \Theta; \sum_{h=0}^{H_n} \left|\upsilon_h\right| \leq \Delta_n, \sum_{h=1}^{H_n} \sum_{k=0}^{d} \left|\omega_{kd}\right| \leq H_n \Delta_n\right\} \tag{14}$$

For given sample size n, we consider only network functions in

$$O_n = O\left(H_n, \Delta_n\right) = \left\{f_{H_n}(x, \theta); \theta \in \Theta\left(H_n, \Delta_n\right)\right\} \tag{15}$$

as an estimate for $m(x)$. Therefore, we redefine the parameter estimate as

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta_n} \sum_{s} \left(y_i - f_H(x; \theta)\right)^2 \tag{16}$$

and the network estimate for $m(x)$ is therefore given by

$$\hat{m}_n(x) = f_{H_n}\left(x, \hat{\theta}_n\right) \tag{17}$$

which is a kind of sieve estimate in the sense of [24] or [25].

To prove consistency of $\hat{T}_{NN}$, it needs to be shown that the Neural Network Based Regression Function $\hat{m}_{NN}$ is also consistent.

Theorem 3.1. *Let* $\left(y_1, x_1\right), \cdots, \left(y_n, x_n\right)$ *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* (12). *Let* $O_n = O\left(H_n, \Delta_n\right), n \geq 1$ *be the set of neural network output functions given by Equation* (15) *with an activation function* ψ *which is Lipschitz continuous on* $\mathbb{R}$, *strictly increasing and satisfying Equation* (13). *Let* $\hat{m}_n(x) = E\left(y_i \mid x_i\right) = x$ *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 the* x_i. *Then* $\hat{m}_n(x)$ *is a consistent estimate of* $m(x)$ *in the* $L^2(F)$*-sense, that is*

$$\int\left(m(x)-\hat{m}_n(x)\right)^2 \mathrm{d}F(x)\to 0 \quad in\ \ probability \tag{18}$$

provided *that* $H_n,\Delta_n\to\infty$ *such that*

$$\Delta_n = o\left(n^{\frac{1}{4}}\right)$$

$$H_n,\Delta_n^4\log n = o(n) \quad and \quad H_n\log n = o\left(\Delta_n^{\alpha}\right)$$

where α *determine the rate of decrease of the tail of the distribution of the* y_i *by Equation* (12).

Proof. Theorem 1 can be proven exactly as Theorem 2.1 of [26] for stationary processes satisfying an α-mixing condition and also as Theorem 3.1 of [27] 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), Lemma A of [28] [29]. Therefore, the right hand side of Equation (5.1) of [26] changes to

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

Then the proof proceeds exactly as in [19] and results in slightly different condition for the rates of H_n,Δ_n in the independence case.

We remark 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 1 essentially is equivalent to Theorem 3.3 of [19]. We also remark that by Theorem 3.4 of [19], we may determine the parameters H_n,Δ_n which determine the network complexity and therefore the smoothness of the function estimate, adaptively from the data by Cross Validation without changing the consistency of $\hat{m}_n(x)$. For the detail on the proof of these theorems, see the work of [26] [27].

Note that, to prove the consistency of $\hat{T}_{NN}$ we need Equation (13) with a simple mean over the unobserved $x_i, i\in r$ instead of the integral. The following results show that the difference between the integral and the sample mean is negligible.

Theorem 3.2. *Let* $\left((y_1,x_1),\cdots,(y_N,x_N)\right)$ *be i.i.d with 3 for some bounded* $m(x)$. *Let F denote the distribution of* x_i. *Let* $|\psi(u)|\le 1$. *Let* $s=1,\cdots,n$ *be the index set of the observed data and* $r=n+1,\cdots,N$ *the index of unobserved data. Let* $\hat{\theta}_n$ *be defined as in Equation* (13) *with* $\hat{m}_n(x)$ *defined as in Equation* (17) *with* $\hat{m}_n(x)=f_{H_n}\left(x,\hat{\theta}_n\right)$ *denote the estimate of* $m(x)$ *based on the sample* $(y_i,x_i), i\in s$. *Let* $n,N\to\infty$ *such that* $\frac{n}{N}\to\pi(0,1)$ *and let* H_n,Δ_n *satisfy conditions in Theorem 3. Then for* $\delta>0$

$$\begin{aligned}&Pr\left(\left|\frac{1}{N-n}\sum_{j\in r}\left(m(x_j)-\hat{m}_n(x_j)\right)^2\right.\right.\\&\left.\left.-\int\left(m(x_j)-\hat{m}_n(x_j)\right)^2\mathrm{d}F(x)\right|>\delta\,|\,(y_i,x_i), i\in s\right)\le d_1\exp\left\{-d_2\frac{N\delta^2}{\Delta_n^4}\right\}\end{aligned} \tag{19}$$

for all $\delta > 0$ *and all N large enough where* $\mathrm{d}_1, \mathrm{d}_2$ *are some constants independent of* N, n *and* $(y_i, x_i), i \in s$.

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 $O(H_n, \Delta_n)$, we immediately have

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

setting

$$V_{N_i} = \left(m(x_j) - \hat{m}_n(x_j)\right)^2 - \int \left(m(x_j) - \hat{m}_n(x_j)\right)^2 \mathrm{d}F(x), i \to r \tag{20}$$

these therefore result to

$$\begin{aligned} &|V_{N_i}| \le 4\left(C^2 + \Delta_n^2\right) \\ &E\left\{V_{N_i} \mid (y_i, x_i), i \in s\right\} = 0 \\ &E\left\{V_{N_i}^2 \mid (y_i, x_i), i \in s\right\} \le 32\left(C^4 + \Delta_n^4\right) \end{aligned} \tag{21}$$

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 apply Bernstein's inequality (Lemma A, Section 2.5.4) of [28] and get

$$\begin{aligned} &Pr\left(\frac{1}{N-n}\left|\sum_{j \in r} V_{N_j}\right| > \delta \mid (y_i, x_i), i \in s\right) \\ &\le 2\exp\left\{-\frac{N_n \delta^2}{64\left(C^4 + \Delta_n^4\right) + \frac{2}{3}4\left(C^2 + \Delta_n^2\right)\delta}\right\} \end{aligned} \tag{22}$$

Now the results follow as $\Delta_n \to \infty$ 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 $\Delta_n = o\left(n^{\frac{1}{4}}\right), \frac{N}{\Delta_n^4} \to \infty$, that is, the right hand side of the inequality converges to zero (taking limits as $\Delta_n \to \infty$).

3.2. Asymptotic Consistency

Theorem 3.3. *If (A1)-(A8) are satisfied and if the activation function* $\psi(u)$ *is Lipschits continuous and strictly increasing and also Theorem* 1 *holds, then the neural network estimate* $\hat{T}_{NN}$ *of the population total T given by 6 with* $\hat{m}_n(x) = f(x, \hat{\theta}_n)$ *and* $\hat{\theta}_n$ *given by* [8] *is consistent in the following sense.*

$$\begin{aligned} &\frac{1}{N}\left|T - \hat{T}_{NN}\right| \to 0 \textit{ in probability} \\ &\textit{where } N, n \to \infty \textit{ with } \frac{n}{N} \to \pi \in (0,1) \end{aligned} \tag{23}$$

provided that the number H_n *and the bound* Δ_n *of the network weights satisfy* $H_n, \Delta_n \to \infty$ *such that*

$$\begin{aligned} &\Delta_n = o\left(n^{\frac{1}{4}}\right) \\ &H_n \Delta_n^4 \log n = o(n) \\ &H_n \log n = o\left(\Delta_n^\alpha\right) \end{aligned} \tag{24}$$

where α *determines* $($*by A1*$)$ *how fast the tail probability of the* ε_i *and* y_i *decreases.* [19] *showed that, the appropriate choice for* Δ_n *is such that* $\Delta_n \to \infty$ *as* $n \to \infty$ *and* $\Delta_n = o\left(n^{\frac{1}{4}}\right)$, *i.e.* $n^{\frac{1}{4}}\Delta_n \to 0$ *as* $n \to \infty$

Proof.

$$\begin{aligned}\frac{1}{N}\left|T-\hat{T}_{NN}\right| &= \frac{1}{N}\left|\sum_{j\in r}\left(y_j-\hat{m}_n\left(x_j\right)\right)\right| \\ &= \frac{1}{N}\left|\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)+\sum_{j\in r}\varepsilon_j\right| \\ &\le \frac{1}{N}\left|m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right|+\frac{N-n}{N}\left|\frac{1}{N-n}\sum_{j\in r}\varepsilon_j\right| \\ & \frac{1}{N}\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)^2+\frac{N-n}{N}\left|\frac{1}{N-n}\sum_{j\in r}\varepsilon_j\right|\end{aligned} \tag{25}$$

by Jensen's inequality.

Now the last term converges to

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

where $(1-\pi)\left|E\left(\varepsilon_j\right)\right| = 0$ since $E\left(\varepsilon_j\right)=0$ by law of large numbers. The first term of 25 decomposes into

$$\begin{aligned}&\frac{N-n}{N}\left(\frac{1}{N-n}\sum_{j\in r}\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)^2-\int\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)^2\mathrm{d}F(x)\right) \\ &+\frac{N-n}{N}\int\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)^2\mathrm{d}F(x)\end{aligned} \tag{26}$$

The right hand terms of 26 converge to 0 by Theorem 1 and as $\frac{N-n}{N}\to 1-\pi$.

The proof is completed by using Theorem 2 to cope with left hand terms where we drop the factor $\frac{N-n}{N}$ converges to $1-\pi$ anyhow.

$$\begin{aligned}&Pr\left(\left|\frac{1}{N-n}\sum_{j\in r}\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)^2-\int\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)^2\mathrm{d}F(x)\right|>\delta\right) \\ &= E\left\{Pr\left[\left|\frac{1}{N-n}\sum_{j\in r}\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)^2-\int\left(m\left(x_j\right)-\hat{m}_n\left(x_j\right)\right)^2\mathrm{d}F(x)\right|>\delta \mid \left(y_i,x_i,i\in s\right)\right]\right\} \\ &\le d_1\exp\left\{-d_2\frac{N\delta}{\Delta_n^4}\right\}\to 0\ \forall\ \delta>0 \text{ as } n\to\infty, \Delta_n\to\infty\end{aligned} \tag{27}$$

hence the proof.

3.3. Mean Squared Error

Mean Squared Error is used to measure the accuracy of the estimator among other measures of performance. The MSE is defined by $E\left(T_{NN}-T\right)^2$ where T denotes the true population total. To estimate $E\left(T_{NN}-T\right)^2$, first, we consider

$$
\begin{aligned}
&E\left[\left(T_{NN}-T\right)^{2} \mid D, X_{n+1}^{N}\right]=E\left[\left(\sum_{i=1}^{H}\sum_{j=n+1}^{N}\hat{m}\left(x,\theta\right)-\sum_{j=n+1}^{N}\left(m\left(x\right)+\varepsilon\right)\right)^{2} \mid D, X_{n+1}^{N}\right]\\
&=\frac{\left(N-n\right)^{2}}{N^{2}}E\left[\left(\frac{1}{N-n}\sum_{i=1}^{H}\sum_{j=n+1}^{N}\hat{m}\left(x,\theta\right)-\sum_{j=n+1}^{N}\left(m\left(x\right)+\varepsilon\right)\right)^{2} \mid D, X_{n+1}^{N}\right]+\frac{N-n}{N}var\left(\varepsilon_{i}\right)\\
&=\frac{\left(N-n\right)^{2}}{N^{2}}E\left[\left(\frac{1}{H\left(N-n\right)}\sum_{i=1}^{H}\sum_{j=n+1}^{N}\hat{m}\left(x,\theta\right)-E\left(T_{k} \mid D, X_{j}\right)+E\left(T_{k} \mid D, X_{j}\right)-E\left(T_{k}\right)\right)^{2}\right]+\frac{N-n}{N}var\left(\varepsilon_{i}\right)\\
&=\frac{\tau_{D}^{2}}{H}\left(1-f\right)\left\{E\left(T_{k} \mid D, X_{j}\right)-E\left(T_{k}\right)\right\}^{2}+\frac{1-f}{N}var\left(\varepsilon_{i}\right)
\end{aligned}
\tag{28}
$$

where the $X_{j=\left(x_{n+1},\cdots,x_{N}\right)}$ is a set of unsampled auxiliary units. T_k denotes the total of the unsampled elements and $E\left(T_k\right)=\sum_{j=n+1}^{N}m\left(x\right)$.

The last approximation of Equation (28) follows from Equation (15) of [30], that is

$$E\left(\frac{1}{HN}\sum_{i=1}^{H}\sum_{j=n+1}^{N}\hat{m}\left(x,\theta\right)-\left(1-f\right)E\left(T_{k}\right) \mid D, X_{j}\right)^{2}\approx\frac{\tau_{D}^{2}}{H}$$

for some positive constant τ_D^2.

The term $E\left(T_k \mid D, X_j\right)-E\left(T_k\right)$ is the predictor bias due to randomness or sampling bias of *D*. Now from Equation (28), we have

$$E\left(T_{NN}-T\right)^{2}=E\left(\frac{\hat{\tau}_{D}^{2}}{H}\right)+\left(1-f\right)^{2}E\left\{E\left(T_{k} \mid D, X_{j}-E\left(T_{k}\right)\right)\right\}^{2}+\frac{1-f}{N}var\left(\varepsilon_{i}\right)\tag{29}$$

As noted in [30], the quantity τ_D^2 can be estimated by batch method. Therefore,

$$\hat{\tau}_{D}^{2}=\frac{s}{r-1}\sum_{t=1}^{r}\left(\hat{T}_{NN,t}-T_{NN}\right)^{2}\tag{30}$$

for details see [30]. Equation (30) can be substituted in 29 in lieu of $E\left(\tau_D^2\right)$.

Now, under the assumption that the $\frac{\varepsilon_i}{\sigma}\sim t\left(v\right)$, then the estimate of $var\left(\varepsilon_i\right)$ is given as

$$v\hat{a}r\left(\varepsilon_{i}\right)=\frac{v}{v-2}\frac{1}{H}\sum_{i=1}^{H}\hat{\sigma}_{i}^{2}\tag{31}$$

Under the assumption that the population is made up of exact copies of the sampled (training) data, we have $E\left(T_k \mid D, X_j\right)-E\left(T_k\right)\approxeq\hat{T}-T$ where $\hat{T}$ the fitted sample totals and

$$E\left(\hat{T}-T\right)^{2}=\left(\sum_{i=1}^{n}\hat{\varepsilon}_{i}\right)^{2}=Var\left(\hat{\varepsilon}_{i}\right)\tag{32}$$

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

$$B\hat{i}as^{2}=\frac{1}{n}v\hat{a}r\left(\varepsilon_{i}\right)\tag{33}$$

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

$$\begin{aligned}\hat{E}\left(T_{NN}-T\right)^2 &= \frac{\hat{\tau}_D^2}{H}+\left(1-f\right)B\hat{i}as^2+\frac{1-f}{N}v\hat{a}r\left(\varepsilon_i\right)\\ &= \frac{\hat{\tau}_D^2}{H}+\frac{1-f}{n}v\hat{a}r\left(\varepsilon_i\right)\end{aligned} \tag{34}$$

As $H\to\infty$ Equation (34) reduces to

$$\hat{E}\left(T_{NN}-T\right)^2=\frac{1-f}{n}v\hat{a}r\left(\varepsilon_i\right) \tag{35}$$

4. Empirical Results

To illustrate our estimation approach, the following data will be utilized. A population of size 188 will be obtained from the United Nations Development Programme 2020 report. The UN studied the development in 1889 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 fall under medium development and ranked number 143 among the 188 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 is considered.

In order to understand how the proposed estimator compares against other existing non-parametric regression estimators, we compared the performance of our estimator to that of identified estimators based on Multivariate Additive Regression Splines (MARS), Generalized Additive Models (GAM) and Local polynomial (LP) which can handle high dimensional data. We compare the performance of the proposed estimator of the population totals, with $\hat{T}_{LP}$, $\hat{T}_{MARS}$, $\hat{T}_{GAM}$ and $\hat{T}_{SAM}$, using the bias, mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

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 criteria are defined as follows; Bias of a Population total estimator refers to the deviation of the expected value of the estimator from the true Total value. **Table 1** provides the results for performance of the estimators when applied to the data obtained from the United Nations Development Programme 2020 report. All of the population total estimators considered here are biased but comparatively T_{NN} exhibits a smaller bias. T_{NN} can be seen to be a very efficient estimator of the finite population total since it has smaller RMSE, followed closely by T_{LP} and T_{MARS}. T_{GAM} proved to be a very inefficient estimator of all other estimators.

Table 1. 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

The conditional performance of the estimator was done and compared with the performance of other existing population total estimators. To do this, 500 random samples, all of sizes 100 and 50, were selected and the mean of the auxiliary values xi 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 were 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 1 and Figure 2 show the behavior of the conditional biases, relative absolute biases and mean squared error realized by all the estimators based on the real data set.

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 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.

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 functions.

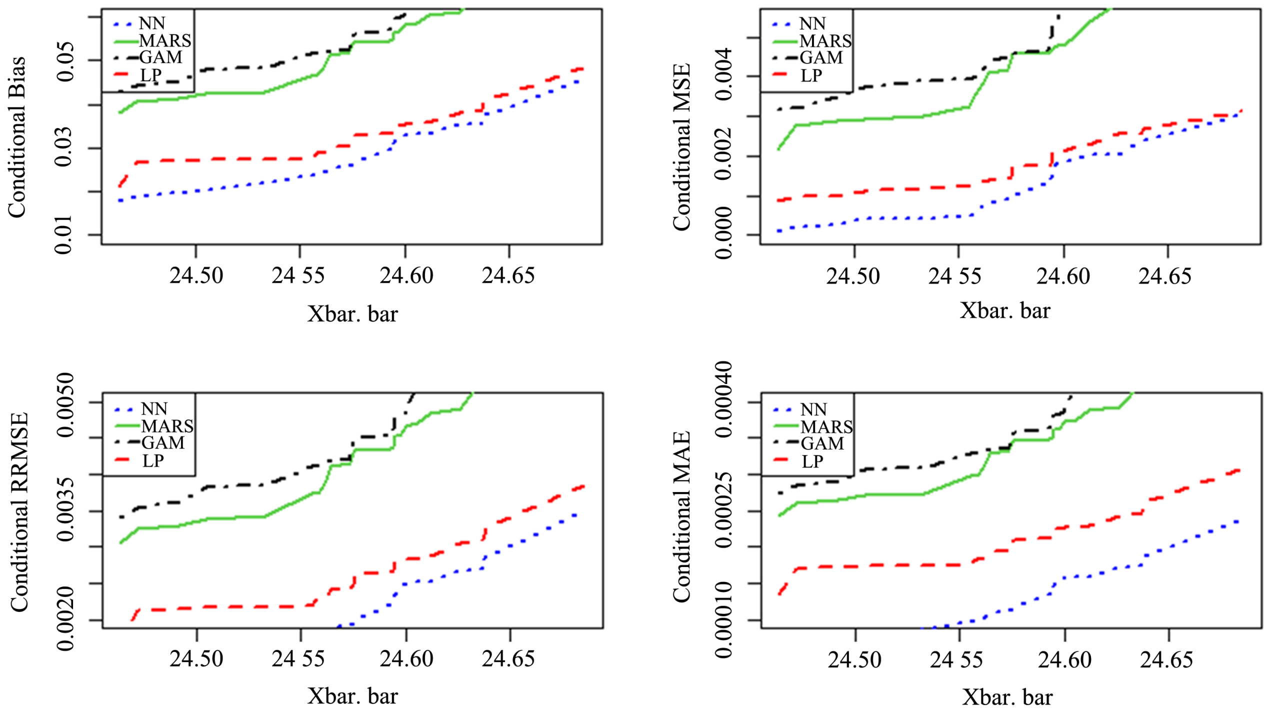

Figure 1. Conditional bias, mean square error, relative root mean square error and mean absolute error based on real data with a sample size of 100.

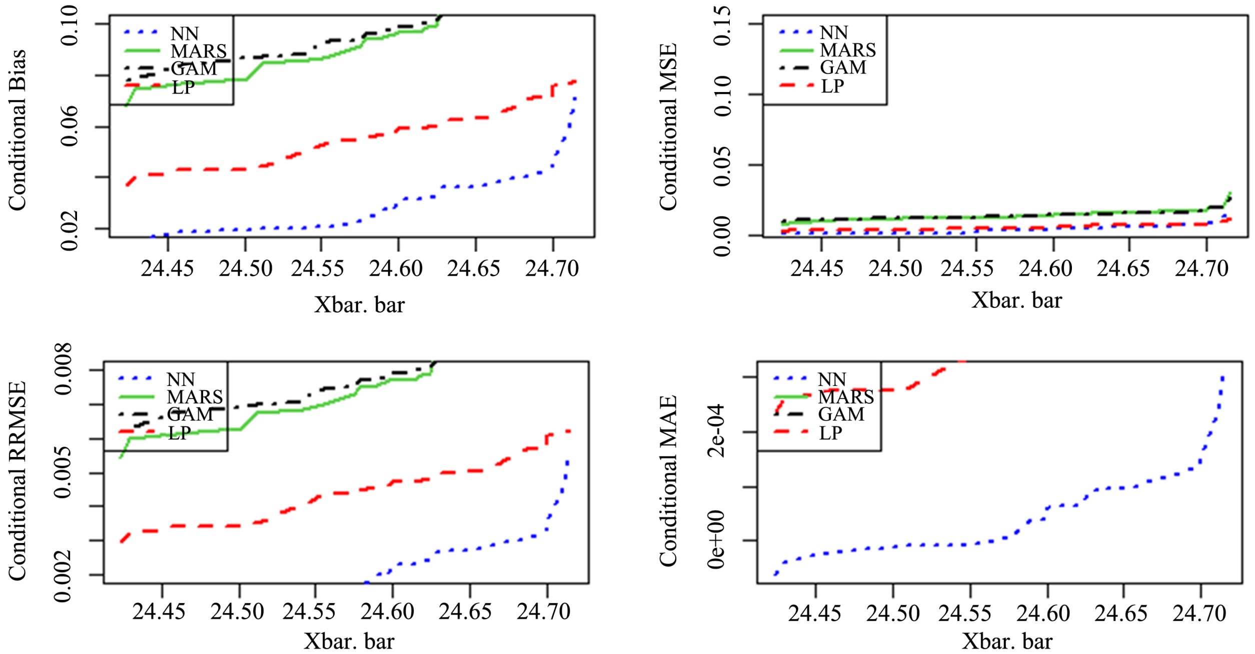

Figure 2. Conditional bias, mean square error, relative root mean square error and mean absolute error based on real data with a sample size of 50.

5. Conclusion and Recommendations

In this paper, an estimator for Finite Population Total has been developed by employing a Feed Forward Back Propagation Neural Network technique in

Non-parametric Regression. Asymptotic properties such as the Consistency and Mean Squared Error for the developed estimator have also been derived. When applied to 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. The developed estimator is considered to be effective in addressing the curse of dimensionality that makes Local Polynomials and Kernel Estimators ineffective when dealing with High Dimensional Data. It should be noted that the proposed estimator has been considered in the case of Simple Random Sampling Without Replacement (SRSWoR). An extension to other sampling techniques such Stratification may be done since they rely on SRSWoR, and it is hypothesised that efficiency will be improved compared to other existing estimators in literature.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Chambers, R.L. and Dunstan, R. (1986) Estimating Distribution Functions from Survey Data. *Biometrika*, **73**, 597-604. https://doi.org/10.1093/biomet/73.3.597

[2] Wang, S.J. and Dorfman, A.H. (1996) A New Estimator for the Finite Population Distribution Function. *Biometrika*, **83**, 639-652. https://doi.org/10.1093/biomet/83.3.639

[3] Hansen, M.H., *et al.* (1987) Some History and Reminiscences on Survey Sampling. *Statistical Science*, **2**, 180-190. https://doi.org/10.1214/ss/1177013352

[4] Dorfman, A.H. (1992) Nonparametric Regression for Estimating Totals in Finite Populations. In: *Proceedings of the Section on Survey Research Methods*, American Statistical Association Alexandria, 622-625.

[5] Otieno, R.O. and Mwalili, T.M. (2000) Nonparametric Regression Method for Estimating the Error Variance in Unistage Sampling.

[6] Jay Breidt, F. and Opsomer, J.D. (2000) Local Polynomial Regression Estimators in Survey Sampling. *Annals of Statistics*, **28**, 1026-1053. https://doi.org/10.1214/aos/1015956706

[7] Hardle, W. and Linton, O. (1994) Applied Nonparametric Methods. In: Engle, R.F. and McFadden, D., Eds., *Handbook of Econometrics*, Vol. 4, Elsevier, Amsterdam, 2295-2339. https://doi.org/10.1016/S1573-4412(05)80007-8

[8] Chambers, R.L., Dorfman, A.H. and Hall, P. (1992) Properties of Estimators of the Finite Population Distribution Function. *Biometrika*, **79**, 577-582. https://doi.org/10.1093/biomet/79.3.577

[9] Montanari, G.E. and Ranalli, M.G. (2003) On Calibration Methods for Design Based Finite Population Inferences. *Bulletin of the International Statistical Institute*, **60**, 2 p.

[10] Stone, C.J. (1982) Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, **10**, 1040-1053. https://doi.org/10.1214/aos/1176345969

[11] Bickel, P.J. and Li, B. (2007) Local Polynomial Regression on Unknown Manifolds. Institute of Mathematical Statistics, Beachwood, Lecture Notes—Monograph Series,

177-186. https://doi.org/10.1214/074921707000000148

[12] Friedman, J.H. (1991) Multivariate Adaptive Regression Splines. *The Annals of Statistics*, **19**, 1-67. https://doi.org/10.1214/aos/1176347963

[13] Di Ciaccio, A. and Montanari, G.E. (2001) A Nonparametric Regression Estimator of a Finite Population Mean. In: *Book Short Papers CLADAG*, Istituto di Statistica, Università degli Studi di Palermo, Palermo, 173-176.

[14] Opsomer, J.D., Jay Breidt, F., Moisen, G.G. and Kauermann, G. (2007) Model Assisted Estimation of Forest Resources with Generalized Additive Models. *Journal of the American Statistical Association*, **102**, 400-409. https://doi.org/10.1198/016214506000001491

[15] El-Housseiny, A.R. and Ziedan, D. (2014) Estimation of Population Total Using Nonparametric Regression Models. *Advances and Applications in Statistics*, **39**, 37-59.

[16] Barron, A.R. (1993) Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory*, **39**, 930-945. https://doi.org/10.1109/18.256500

[17] Ken-Ichi, F. (1989) On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks*, **2**, 183-192. https://doi.org/10.1016/0893-6080(89)90003-8

[18] Cybenko, G. (1989) Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, **2**, 303-314. https://doi.org/10.1007/BF02551274

[19] White, H. (1990) Connectionist Nonparametric Regression: Multilayer Feed forward Networks Can Learn Arbitrary Mappings. *Neural Networks*, **3**, 535-549. https://doi.org/10.1016/0893-6080(90)90004-5

[20] Franke, J. and Neumann, M.H. (2000) Bootstrapping Neural Networks. *Neural Computation*, **12**, 1929-1949. https://doi.org/10.1162/089976600300015204

[21] Gene Hwang, J.T. and Ding, A.A. (1997) Prediction Intervals for Artificial Neural Networks. *Journal of the American Statistical Association*, **92**, 748-757. https://doi.org/10.1080/01621459.1997.10474027

[22] Barron, A.R. (1994) Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning*, **14**, 115-133. https://doi.org/10.1007/BF00993164

[23] Asnaashari, A., McBean, E.A., Gharabaghi, B. and Tutt, D. (2013) Forecasting Watermain Failure Using Artificial Neural Network Modelling. *Canadian Water Resources Journal*, **38**, 24-33. https://doi.org/10.1080/07011784.2013.774153

[24] Grenander, U. and Ulf, G. (1981) Abstract Inference. Technical Report.

[25] Geman, S. and Hwang, C.-R. (1982) Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *The Annals of Statistics*, **10**, 401-414. https://doi.org/10.1214/aos/1176345782

[26] Franke, J. and Diagne, M. (2006) Estimating Market Risk with Neural Networks. *Statistics & Decisions*, **24**, 233-253. https://doi.org/10.1524/stnd.2006.24.2.233

[27] Shen, X.X., Jiang, C., Sakhanenko, L. and Lu, Q. (2019) Asymptotic Properties of Neural Network Sieve Estimators.

[28] Serfling, R.J. (1980, 2000) Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Inc., Hoboken. https://doi.org/10.1002/9780470316481

[29] Serfling, R.J. (2009) Approximation Theorems of Mathematical Statistics, Volume 162. John Wiley & Sons, Hoboken.

[30] Liang, F.M. and Kuk, Y.C.A. (2004) A Finite Population Estimation Study with Bayesian Neural Networks. *Survey Methodology*, **30**, 219-234.