International Journal of Data Science and Analysis
2020; 6(1): 1-11
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20200601.11
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Estimating Total Energy Demand from Incomplete Data Using Non-parametric Analysis

Benard Mworia Warutumo*, Pius Nderitu Kihara, Levi Mbugua

Department of Statistics and Actuarial Science, Technical University of Kenya, Nairobi, Kenya

Email address:
moriahbm123@gmail.com (B. M. Warutumo)

*Corresponding author

To cite this article:
Benard Mworia Warutumo, Pius Nderitu Kihara, Levi Mbugua. Estimating Total Energy Demand from Incomplete Data Using Non-parametric Analysis. *International Journal of Data Science and Analysis*. Vol. 6, No. 1, 2020, pp. 1-11. doi: 10.11648/j.ijdsa.20200601.11

Received: November 12, 2019; **Accepted**: December 6, 2019; **Published**: January 8, 2020

Abstract: The validity and usefulness of empirical data requires that the data analyst ascertains the cleanliness of the collected data before any statistical analysis commence. In this study, petroleum demand data for a period of 24 hours was collected from 1515 households in 10 clusters. The primary sampling units were stratified into three economic classes of which 50% were drawn from low class, 28% from medium class and 22% from high class. 63.6% of the questionnaires were completed whereas incomplete data was computed using multivariate imputation by chained equation with the aid of auxiliary information from past survey. The proportion of missing data and its pattern was ascertained. The study assumed that missing data was at random. Nonparametric methods namely Nadaraya Watson, Local Polynomial and a design estimator Horvitz Thompson were fitted to aid in the estimation of the total demand for petroleum which has no close substitute. The performance of the three estimators were compared and the study found that the Local Polynomial approach appeared to be more efficient and competitive with low bias. Local polynomial estimator took care of the boundary bias better as compared to Nadaraya Watson and Horvitz Thompson estimators. The results were used to estimate the long time gaps in petroleum demand in Nairobi county, Kenya.

Keywords: Clean Data, Missing Data, Imputation, Petroleum Total Demand

1. Introduction

Data is a long lived asset used in many unforeseen ways including data mining, mass customization and optimization. Data need to be of high quality so that decisions can be made on the basis of its reliability and validity. Quality data is an accurate representation of the part of the "real world" that it models and should be fit for the purpose of which it is designed for [1]. This entails that data ready for analysis should be accurate, precise, complete, current, non-redundant, portable and credible. Poor quality data is costly to diagnose and repair with most costs being hidden and hard to quantify. More specifically, incomplete data leads to extra resources in correcting, dealing with reported errors and reworking. The consequences are adverse across disciplines from loss of revenue through customer dissatisfaction in business, lowered employee morale in human resource among others. It is thus imperative that to improve on the quality of data, appropriate tools and techniques need to be developed [2]. Most data quality problems emanate from poorly defined concepts, incomplete questionnaires, inaccurate data entry and checking processes [3]. To mitigate these problems, a robust data quality plan is eminent and should include every definition of a record, its timeliness, completeness, accuracy and how it will be monitored. With complete data statistical inference can provide a transitional framework to generate insight to inform decision. Thus identifying any data gaps before other factors are considered is imperative.

In the energy sector, to facilitate the installation of different energy storage mechanisms including pump dispensers, pipe work at service stations, consumer installations and highly inflammable gas requiring huge safety and environmental implications data driven informed decisions help to avoid calamities as evidenced with oil spill over [4]. In most developing countries, analysis of energy demand has concentrated on innovative ways of switching from the use of non-renewable energy to renewable energy like wind substituting diesel as a source of power for pumping, converting diesel powered process machinery to electrically

powered machines, briquetting agricultural residuals to replace kerosene as cooking fuel in urban areas, alcohol for transport vehicles to replace diesel is another thought among others. Thus with discovery of oil in some developing countries and depletion of oil in others, modeling energy demand has become essential due to its adverse consequences in terms of environmental and its overall cost implication [5].

2. Incomplete Data due to Missing Values

Missing values can either be omitted at random (MAR) or omitted not at random (MNAR). Omitted at Random implies that the propensity for a data point to be missing is not related to the omitted data, but it is related to some of the observed data. This has nothing to do with the missing values but has to do with the values of some other variable. On the other hand, for MNAR, the probability of a value missing varies for reasons that may be unknown. To distinguish between MNAR and MAR it is imperative to assess the nature of data and its quality in terms of completeness, the use of descriptive statistics such as frequency polygons as well as the scatter plots and more advanced methods including the use of estimators such as Nadaraya-Watson, Horvitz Thompson and Local Polynomial are some of the most promising methods of solving data gaps [6]. For empirical data analysis, where the samples are identical and independently distributed, non-parametric estimation procedure with a bandwidth parameter can be used as a modeling technique for the missing values. The starting point is identifying a sample S of n paired observations $(x_i, y_i), i = 1,2, \dots, n$ from a population$U$, of size N. This enables one to find an estimator for $E(y_i) = g(x_i)$ of a missing population [6].

2.1. Multiple Imputation for Missing Data

In survey, missing values as a result of non-response can be addressed through multiple imputation (MI) methods. In singular imputation methods, a measure of central tendency is used to input the missing values. Multiple imputation narrows uncertainty about missing values by creating several different imputation options of which several forms of the same data set are created, which are then combined to make the most optimal values. When used correctly MI can reduce bias, improve validity, increase precision and the results are robust and resistant to outliers [7]. Missing data point can be estimated by the average values of the parameter estimate obtained as point estimate based on the standard errors. This involve combining prior information about a parameter of interest with new evidence from a sample or through resampling with an appropriate probability distribution [7]. A naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests [8].

2.2. The Chained Equation Approach to Multiple Imputation

Multivariate imputation by chained equations (MICE) operates under the assumption that missing data is at Random (MAR), which means that the probability that a value is missing depends only on the observed values [9]. Many of the initially developed multiple imputation procedures assume a large joint model for all the variables, such as a joint normal distribution of which for large datasets, with hundreds of variables of varying types, this is rarely appropriate [7]. Multivariate imputation by chained equations (MICE) is a flexible approach in which a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable can be modeled according to its distribution whereby binary variables are modeled using logistic regression, continuous variables modeled using linear regression, multinomial logit model for categorical variables and a Poisson model for count variables. In MICE procedure, all variables that are to be used in subsequent analyses, whether or not they have missing data may be predictive of the missing values and are included in the imputation process. One key point is to include the variables that are likely to satisfy the MAR assumption. Beyond that, the specific issues that often come up when selecting variables include: creating an imputation model that is more general than the analysis model; imputing variables at the item level vis a vis the summary level and imputing variables that reflect raw scores vis a vis standardized scores.

3. Nonparametric Analysis

Non-parametric methods estimate the distance between a point and its neighbors and the estimation depends heavily on the bandwidth and its span. These methods consider correlation between available auxiliary data-set and the missing response variable as missing at random. The general nonparametric regression models are either of fixed or of random design, such that if n data points $(x_1, y_1), (x_2, y_2). \dots (x_n, y_n)$ have been collected, the relationship is modeled as

$$y_i = mX + e_i \tag{1}$$

Where, X is the predictor variable also known as the regressor, Y is the response variable and m is the unknown regression function with observation error e_i of which$E(e) = 0$ and$var(e) = \sigma^2$. The fixed design model is concerned with controlled non-stochastic repressor X variable, this implies that the regressors are controlled by the researcher and are simply assumed to be measured without error. In fixed design, for any given observation (x_i, y_i), $x_i \in \Re^d$ and Y is an independent random variable with $E(Y) = m(X)$. Random design models are used in observational studies and are common in non-experimental science. The observed predictor variables are independent and identically distributed (iid) random variables such that

$$m(x) = E(Y|X) = \int_y y \frac{f(x,y)}{f(x)} \, \partial y \tag{2}$$

Where $f(x, y)$ is the joint density of (X, Y) and $f(x) = f_X(x)$ is the marginal probability function of x. The approximation of the function $m(x)$ is through smoothing method of the form $\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} w_{ij} Y_{ij}$, where w_{ij} are the

weights. The common smoothing methods include the kernel, the nearest neighbor, the orthogonal series and the spline method. This study concentrated on kernel techniques which are linear estimators in that the value of the estimator at any point x is the weighted sum of the responses. The weight function is defined as:

$$K_h(x - x_i) = \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \quad (3)$$

With the function K supported on $[-1,1]$ that has a maximum at zero. The shape of the weights satisfies the moment conditions

$$\int_{-1}^{1} K(u)du = 1\,,\ \int_{-1}^{1} uK(u)du = 0$$

$$\int_{-1}^{1} u^2 K(u)du \neq 0 \ and\ V = \int_{-1}^{1} K(u)^2 du < \infty$$

Where K is symmetric about zero. The kernel with finite support can be rescaled to have support on $[-1,1]$. Kernels, which have infinite support on the entire line, result to estimators with global bias difficulties. The smoothing parameter h determines the size of the weights. Small h leads to wigglier (rougher) estimators while larger h leads to a more averaging (horizontal) estimator.

3.1. Nadaraya and Watson Estimator

Assume the observation of some variable Y have been taken n times for some utility at times $t_1 \cdots, t_n$. Let y_i, be decomposed into two parts, $m(\cdot)$ the regression function which represents the true underlying change curve following the economic and physical potential and ε_i the errors which may not depend on time. These errors not only stand for observational error but also for economic random variation due to seasonal and other exogenous factors. We assume that $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq 1$ where $t_1, t_2, \ldots, t_n$ is the explanatory variable analogous to $x_1, x_2, \ldots, x_n$ for ease of notation. A kernel estimator $\hat{m}(t_0)$ for $m(t_0)$ can be written as:

$$\hat{m}(t_0) = \sum_{i=1}^{n} w_i(t_0; t_1 \cdots, t_n: h) y_i$$

where w_i, are the weights given by

$$w_i = (t_0; t_1, \cdots, t_n;\ h) = \sum_{i=1}^{n} K_h(t - t_i) / \sum_{j=1}^{n} K_h\big(t - t_j\big)$$

The weights do not depend on $\{Y_i\}$ and therefore $\hat{m}(t_0)$ is a linear estimator which can be expressed as a minimiser of the locally weighted least squares

$$\hat{m}(t_0) = \sum_{i=1}^{n} \big\{Y_i - \sum_{j=0}^{p} \beta_j (t_1 - t_0)^j\big\}^2 K_h(t_i - t_0) \quad (4)$$

Where the squared part of the right hand side of (4) represents the polynomial part and the other part represents the local constant. While the sum ranges from 1 to n, only those y_i lying in the interval $(t_0 - h, t_0 + h)$ contribute to $\hat{m}(t_0)$. This leads to

$$m_h(x) = \sum_{i=1}^{n} K\big(h^{-1}(x - x_i)\big) y_i / \sum_{i=1}^{n} K\big(h^{-1}(x - x_i)\big) \quad (5)$$

For simplicity, we define (5) as $\hat{g}(t)/\hat{f}(t)$ which are the finite sample approximation to

$g(t) = \int_{-\infty}^{\infty} yf(t, y)dy$ and $f(t) = \int_{-\infty}^{\infty} f(t, y)dy$

3.2. Local Polynomial Estimator

The idea of local polynomial estimators has been widely studied by [10], [11] and [12]. Suppose that Locally the regression function m can be approximated by

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!} (z - x)^j \equiv \sum_{j=0}^{p} \beta_j (z - x)^j \quad (6)$$

For z in a neighborhood of x. By using Taylors expansion, m (z) can be modeled locally by a simple polynomial model. This suggests using a locally weighted polynomial regression of the form

$$\sum_{i=1}^{n} \big\{Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j\big\}^2 K_h(X_i - x) \quad (7)$$

Where K (.) denotes the kernel function and h is a bandwidth as presented in (3). Denote by $\hat{\beta}_j\ (j = 0, \ldots, p)$ the minimizer of the equation (7). The above exposition suggests that the estimator for $m^{(v)}(x)$ is

$$\hat{m}_v(x) = v!\, \hat{\beta}_v \quad (8)$$

Where the whole curve $\hat{m}_v(.)$ is obtained by computing the local polynomial regression with x varying in an appropriate estimation domain. With p=1, the estimator $\hat{m}_0(x)$ is termed a local linear regression smoother or a local linear fit. This estimator can be explicitly expressed as

$$\hat{m}_0(x) = \frac{\sum_1^n w_i Y_i}{\sum_1^n w_i} \quad (9)$$

Local polynomial fitting is an attractive method both from the theoretical and practical point of view [11]. The method adapts to various types of designs including random, fixed design, highly clustered and nearly uniform design. There is an absence of boundary effect, the bias at the boundary stays automatically of the same order as in the interior, without use of specific boundary kernels. The asymptotic minimax efficiency is100% among all linear estimators and only a small loss has to be tolerated beyond this class [11].

3.3. The Horvitz-thompson Estimator of Total

Given a finite population of N individuals, and we are interested in some trait that they have. Let X_j denote the value of the trait for individual j We don't get to see all these $X_j's$ we only sample $n < N$ of them. With this sample of n individuals, we may be interested in obtaining an estimate of the total $T = \sum_{j=1}^{n} X_j$ or the mean

$$\tau = \frac{1}{N} \sum_{j=1}^{N} X_j \quad (10)$$

If the probability of individual j being included in the sample is π_j the Horvitz-Thompson estimator of the total is given as:

$$\hat{T}_{HT} = \sum_{j-1}^{n} \frac{x_j}{\pi_j} \quad (11)$$

This estimator gives each observation a weight which is the inverse of its probability of inclusion and under unequal probability sampling, the Horvitz-Thompson estimator (HT estimator) is an unbiased estimator of the population total [13]. This estimator can also be defined as:

$$\hat{t}_{HT} = \sum_{j-1}^{n} \frac{x_j}{\pi_j} = \sum_{i \in s} w_j X_j \quad (12)$$

where w_j is the design weight of the jth element. The HT estimator of the population mean can be expressed as

$$\hat{\bar{y}}_{HT} = \frac{\hat{t}_{HT}}{\hat{N}} = \frac{1}{\hat{N}} \sum_{j \in s} \frac{y_j}{\pi_j} = \frac{1}{\hat{N}} \sum w_j y_j \quad (13)$$

Where $\hat{N} = \sum_{j \in s} w_j$ is the estimated population size. With the same functional form, we can rescale the weights so that $\sum_{j \in s} w_j = n$ and rewrite (13) as

$$\hat{\bar{y}}_{HT} = \frac{1}{n} \sum_{j \in s} \breve{w}_j y_j$$

The design based Horvitz –Thompson estimator of population total will thus be given by

$$\hat{T}_{HT} = \frac{Y}{n} \sum \frac{m_j \bar{y}_j}{x_j} \quad (14)$$

Where $\bar{y}_j = \frac{\sum_{j \in s} y_{ij}}{n_j}$, $x_j, j = 1,2, \ldots, N$ are known auxiliary variables, further $X = \sum_{j=1}^{N} x_j$. The efficiency of Horvitz-Thompson estimator can be improved by using auxiliary information x_i to model the finite population $y_j's$ as a realizations from an infinite super population model, ξ, relating x_j to y_j via equation (1).

3.4. Model Based Estimator of Population Total

The potential disadvantage of estimators motivated by super population model is inefficiency under model misspecifications [14]. This could be avoided by replacing the parametric specification by a non-parametric specification in which $m(.)$ is smooth function of x_j and $V(.)$ is smooth and strictly positive [4]. To enhance efficiency, the model based estimator of population total using Nadaraya Watson is where, $\varphi(\cdot)$ is chosen particularly to be the form of the bias corrected version of the AIC. $tr(\cdot)$ is the trace of the smoothing matrix regarded as the nonparametric version of degrees of freedom, called the effective number of parameters [5]. When $\varphi(tr(\cdot)) = -2log(1 - tr(S_h)/n)$, then (16) becomes the generalized cross validation criterion.

given by

$$\hat{t}_{Nw} = \sum_{i \notin s} \mu_n(xi) + \sum_{i \in s} y_j \quad (15)$$

Where $\mu_{n(xi)}$ is as defined in (5).

The model based estimator of total using local polynomial is given by

$$\hat{t}_{Lp} = \sum_{i \notin s} \mu_0(xi) + \sum_{i \in s} y_j$$

Where $\mu_0(xi)$ is defined in (9).

The model based estimators are highly efficient when $\mu_n(xi)$ and $V(x_j)$ are correctly specified but biased and even inconsistent if the model is wrong. In surveys involving clusters with equal number of second stage units, equal probability of selection is used [13]. In many surveys however, the second stage samples are not equal. This research presents a case where the Secondary Sampling Units, SSU are not necessarily equal.

3.5. Smoothing Parameter Selection

In nonparametric kernel estimation, the smoothing parameter effectively controls the model complexity. When $h \to \infty$, local modeling becomes a global modeling, when $h = 0$ the estimate essentially interpolates the data and the modeling bias will be small. Since the bias is proportional to h^2 and the variance proportional to $\frac{1}{h}$, the bandwidth has to be taken neither too large nor too small so as not to increase the bias and variance of the estimates [5]. The problem can be solved theoretically by choosing a bandwidth that balances the trade-off between the bias and the variance components, since the consistency of the estimator is based on the sum of the bias and variance.

The positive value h that minimizes any of the selection criteria namely AIC, BIC and AIC_c is selected as an optimal smoothing parameter. In this study, the smoothing parameter adopted was the Improved Akaike Information Criterion (AIC_c), derived from the classical AIC for linear models under the likelihood setting:

$$-2(maximized\ loglikelihood) + 2(number\ of\ estimated\ parameters)$$

Thus, select h minimizing

$$AIC_c = log\frac{\sum\{y_i - \hat{f}_h(x_i)\}^2}{n} + 1 + \frac{2\{tr(S_h)+1\}}{n - tr(S_h) - 2}$$

$$= log\frac{\|(S_h - I)y\|^2}{n} + 1 + \frac{2\{tr(S_h)+1\}}{n - tr(S_h) - 2}$$

$$= log(\hat{\sigma}^2) + \varphi(tr(S_h), n) \quad (16)$$

4. Results and Discussion

In this study, the population of interest comprised of all the households within Nairobi County, Kenya, which was projected at 1,551,06. The sampling frame was constructed

using the Kenya Population Census of 2009 which had 10,323 enumeration areas (EAs). Petroleum demand data in kilogram of oil equivalent (koe) for a period of 24 hours was collected from 1515 households which came from 10 clusters. A total of 963 households comprising of 63.6 per cent had complete interviews, 43 households comprising of 3.3 percent gave partial answers while 457 households (30.2 per cent) of the households could not be accessed. The other data source of auxiliary population characteristics from the clusters in question were obtained from the records of previous survey [18-20].

The total cost of the survey was computed as,

$$C = a \times C_1 + a \times b \times C_2$$

Where a is the primary sampling units and b the additional households [15]. With their unit costs being C_1 and C_2 respectively. Thus the average number of households in each primary sampling unit was computed as:

$$b_{opt} = \sqrt{\frac{C_1(1-P)}{C_2 \times P}} \quad (17)$$

P is the proportion obtained from a pilot survey conducted before the main survey [15].

In this study, C_1 was estimated to be KES 50,000 and C_2 was 2KES. The value of P was given by 84/100, the proportion of persons who used petroleum products during the pilot survey. Thus, b_{opt} was found to be 70 households in each Primary sampling unit (PSU). The number of PSUs used was computed to be 9.97 approximately 10 enumeration areas. Thus a total of (70X9.97) household's interviews were needed to qualify for this survey. To minimize non-response, the sample was inflated by 10% leading to a total sample size at 757 households in 10 PSUs. Random numbers were used to pick 288 clusters from a sampling frame obtained from the National Sample Survey and Evaluation Programme (NASSEP IV) consisting of 10,323 clusters in Nairobi County. Primary sampling units were stratified into low class representing 50 percent, middle economic class representing 28.8 percent and high economic class representing 22 percent. This was done using Probability Proportion Sampling (PPS scheme) taking care of the clustering effect design, see Table 1.

Table 1. *Primary sampling units selected for the study.*

Econmic Class	PSU Serial number	Expected number of	Range	Random number between 1 and range	PSUs selected
Low Class	1,2,…144	5	29	2	2,31,60,88,117
Medium	1,2…81	3	27	8	8,35,62
High	1,2…63	2	31	24	24,55
Total	1,2…288	10	Total Clusters Selected		10

Source: Household petroleum survey conducted in the country of Nairobi-aug-september.

In this study, incomplete data was found to be due to non-response as a result of language barrier, failure to identify sampled respondent and loss of completed questionnaires by some enumerators. An evaluation of proportion of population with missing data and its pattern provided a diagnosing mechanism for a reasonable method of imputing missing values. Figure 1 depicts the proportion of population missing data and its missing pattern.

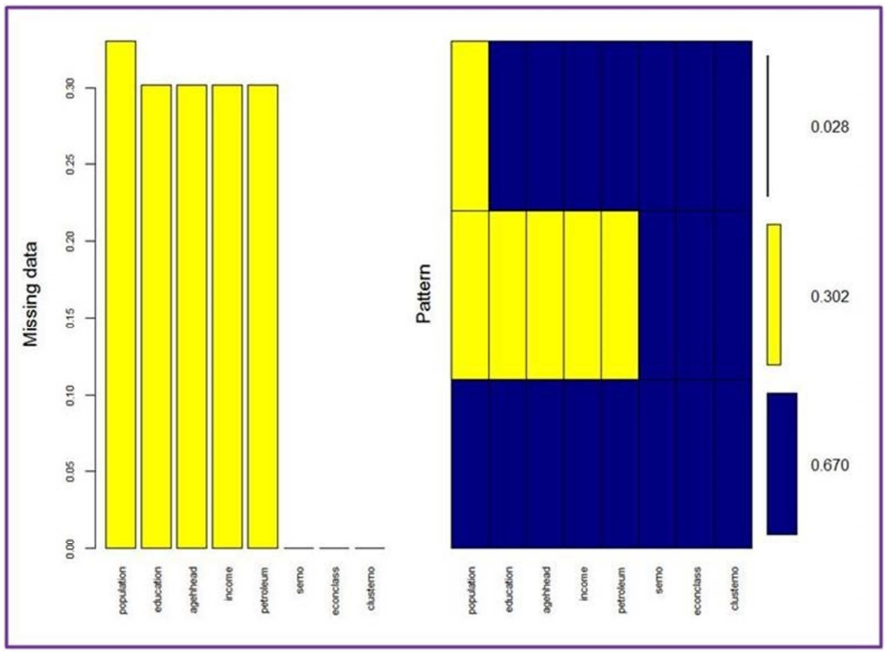

Figure 1. *The Proportion of missing data and the corresponding missing pattern.*

Age and income of household head had the most missing values with 33 percent of the expected respondents not

giving complete information. This implied that we would lose 67 percent of all the petroleum demand information collected within the 1515 households if missed data was not corrected through an imputation process. Missing information was imputed using auxiliary information based on the head of the households' age, income, education status and the household population. This was done using the multivariate imputation by chained equations (MICE) under the assumption that missing data was at random. The observed values and the imputed values for the 10 clusters are presented in Table 2.

Table 2. *Imputed household Petroleum demand obtained through multivariate imputation by chained equations.*

Cluster Number	Observed	Imputed	Total Demand
1.00	62.38	41.31	103.69
2.00	314.72	224.84	539.56
3.00	84.05	70.33	154.38
4.00	164.18	18.89	183.07
5.00	64.48	46.04	110.52
6.00	57.59	71.38	128.97
7.00	376.57	396.82	773.39
8.00	60.94	40.09	101.03
9.00	359.72	206.38	566.10
10.00	297.66	130.73	428.39
Total	1,842.29	1,246.81	3,089.10

Kilograms of oil equivalent (koe).

In estimating the total demand in the population, the model based estimator (15) which uses Nadaraya-Watson, the model based estimator (16) which uses local polynomial, and the HT estimator equation (14) were compared. The bandwidth obtained from AICC was 6.6.

For the Nadaraya-Watson Estimator, the bias concentrated between 165koe and 428koe. This estimate showed sparseness of data between the clusters with population of 450 to 550 persons, which was not also equally spaced.

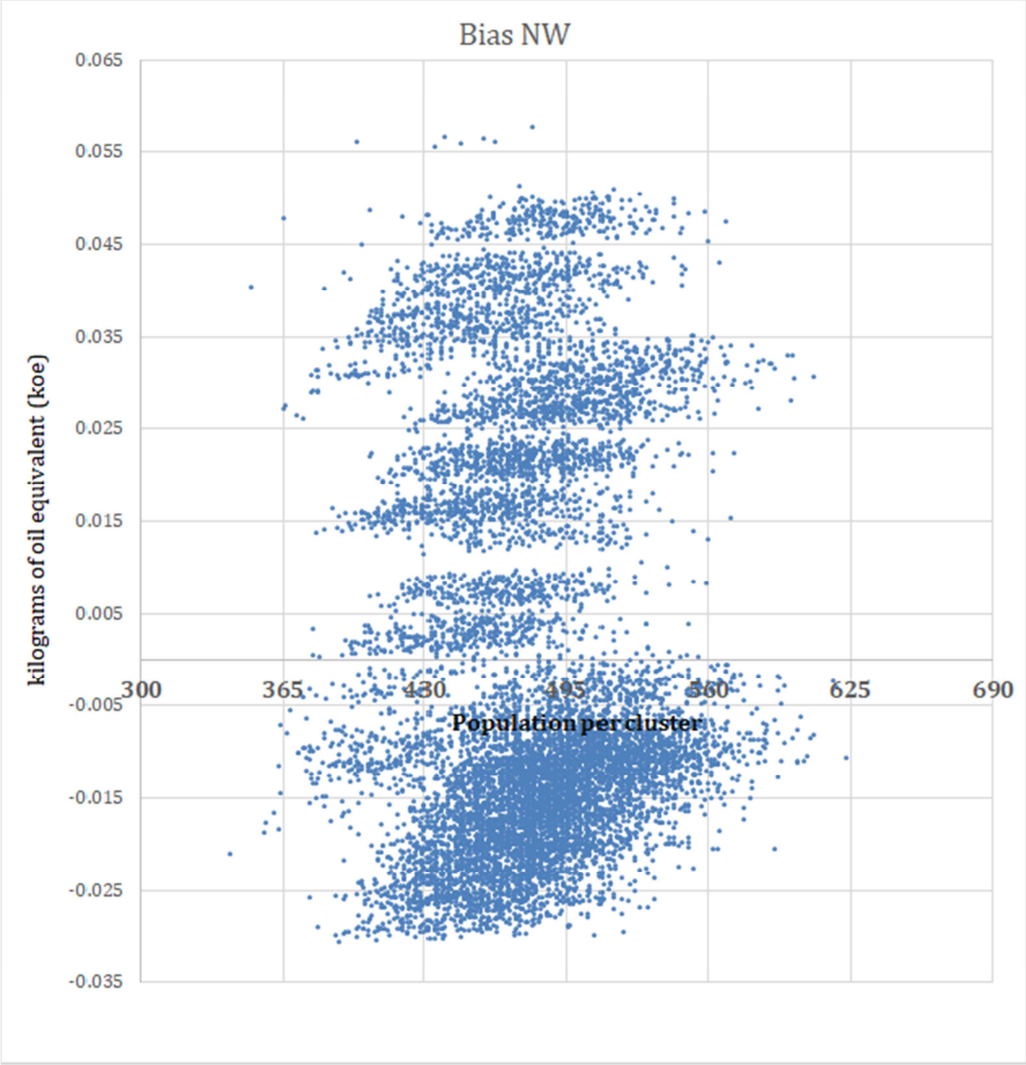

Figure 2. *The bias of Nadaraya Watson Estimator.*

Figure 2 shows that the bias ranged between -0.03049 to 0.05771 with the highest mean square error of 49,926.22koe as shown in Figure 3.

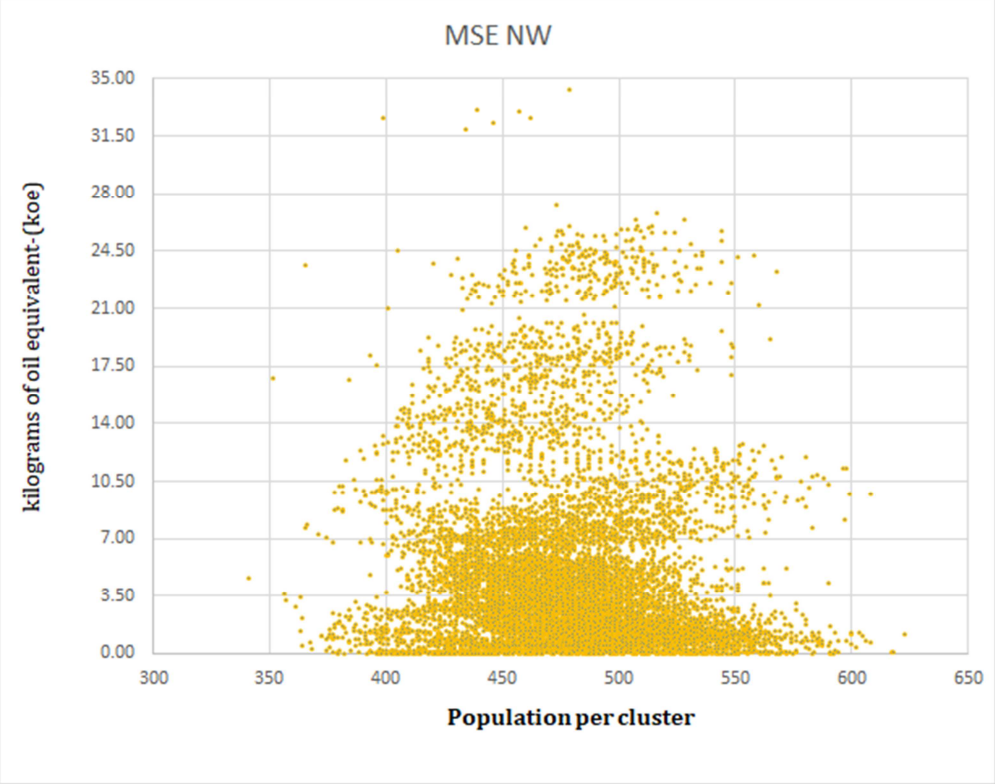

Figure 3. *The Mean Square Error of Nadaraya Watson Estimator.*

The mean square error MSE seems to be in clusters lying between 400 and 550 persons with more than 99% of the demand ranging between 365 and 625. Figure 4 shows the boundary effect of the Nadaraya-Watson estimator with some inconsistency at the boundary and not clearly showing the ten clusters.

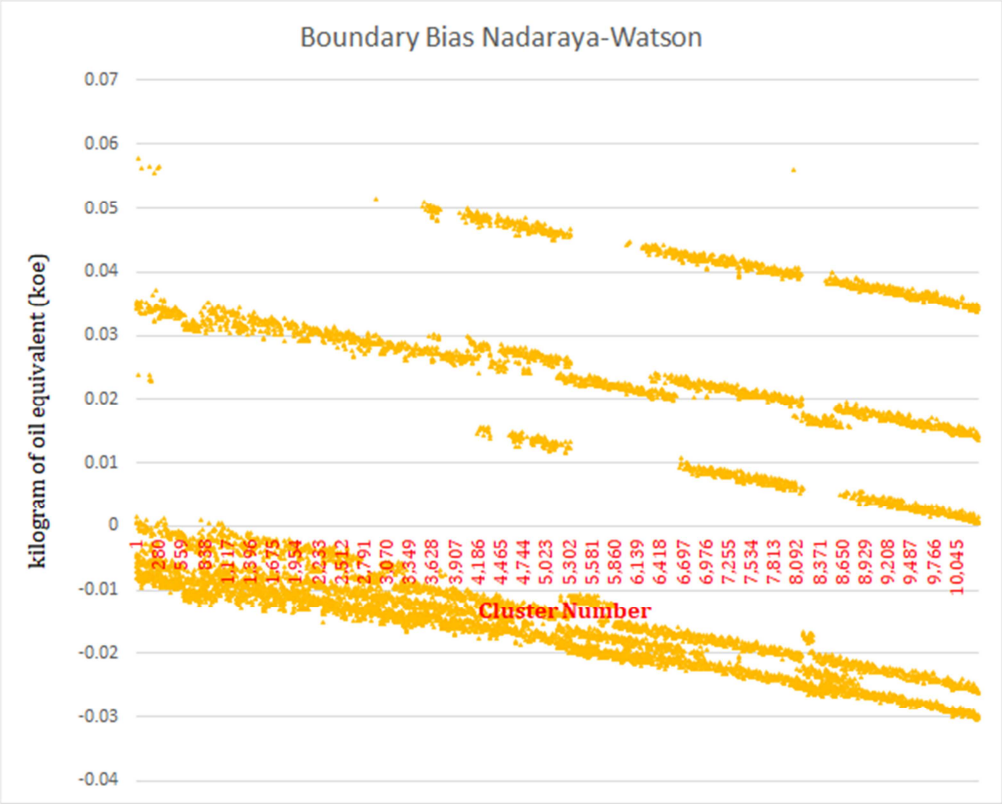

Figure 4. *The Nadaraya Watson boundary bias.*

Overall, Nadaraya-Watson indicated an estimated total demand of 3,061,485.90koe.

For local polynomial model based estimator, the mean squared error reported was 0.10417 koe with the estimate ranging between 101.03 to 940.28 koe. The bias of local polynomial estimation was between -0.00004 to 0.00005 koe with the bulk of its bias lying around the zero mark. The boundary bias captured all the ten clusters unlike the Nadaraya-Watson estimator as shown in Figure 5. The Local polynomial estimator was consistent at the boundary with an estimated total demand of 3,061,485.90koe.

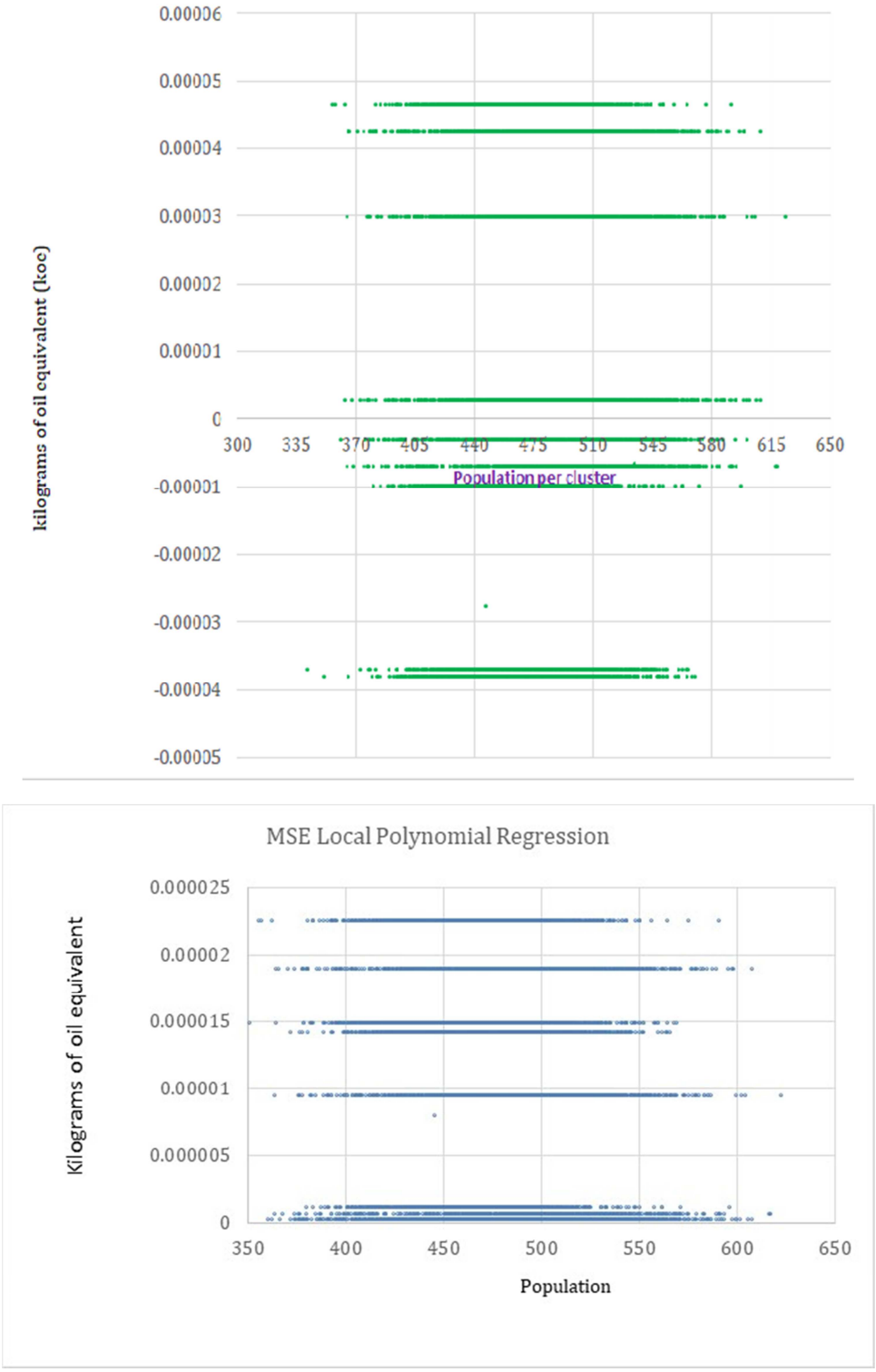

***Figure 5. L**ocal Polynomial boundary bias.*

For the Horvitz Thompson estimator, the average bias ranged -0.02734 to 0.06427 as indicated in Figure 6 while Figure 7

illustrates the boundary bias of Horvitz Thompson Estimator Horvitz Thompson estimator.

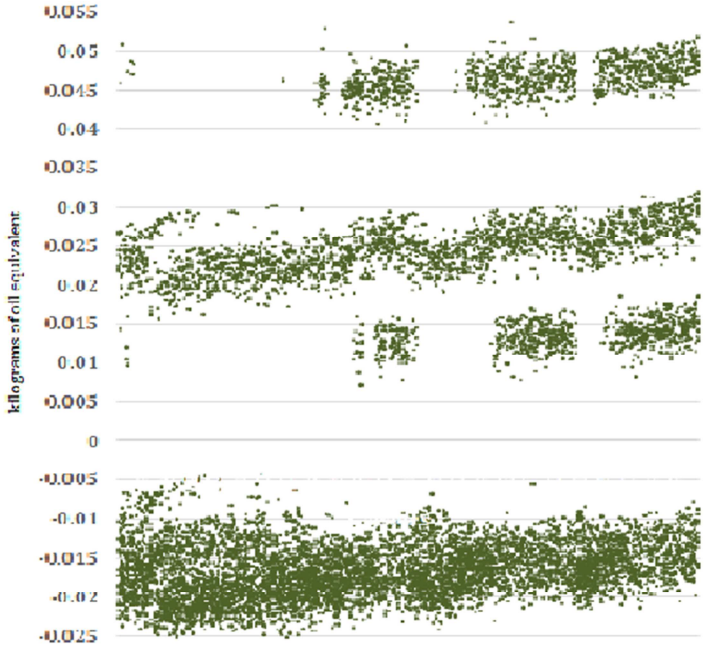

Figure 6. *Horvitz Thompson Estimator bias.*

Of the three estimators, Local polynomial had the lowest mean squared error as compared to Horvitz Thompson Estimator which reported the highest MSE followed by Nadaraya-Watson respectively. Horvitz Thompson estimator bias ranged from -0.02734 to 0.06427koe. This information should come in section. Table 3 reports the total estimated demand for the three classes namely low class, middle class and high class as estimated by the Local Polynomial estimator.

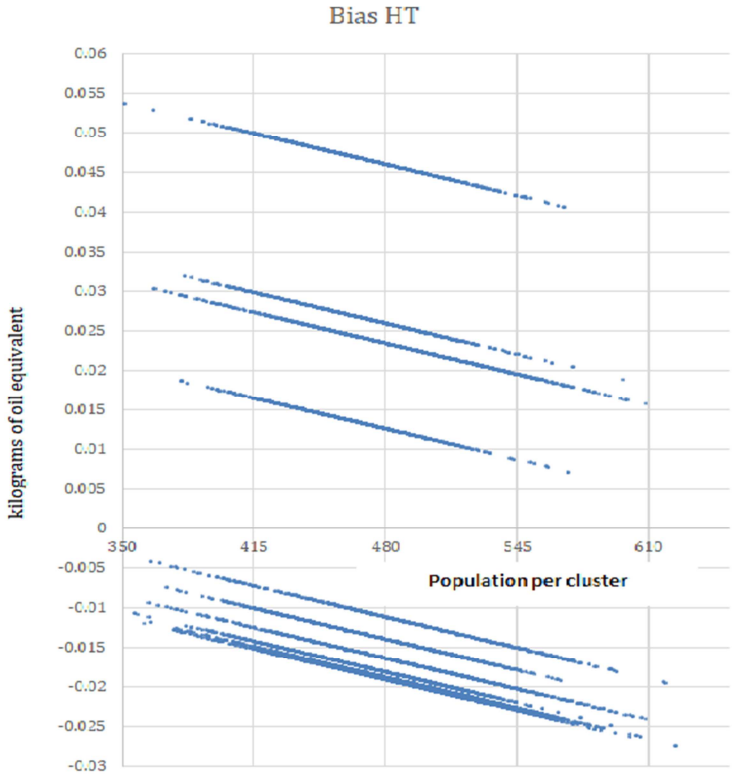

Figure 7. *The boundary bias of Horvitz Thompson Estimator.*

Table 3. *Final Petroleum Demand as estimated by the Local Polynomial Estimators for the Low, Middle and High Classes.*

Summary	Estimator	N	Sum	Minimum	Maximum
Bias (koe)	Nadaraya Watson Estimator	10313	0.00563	-0.03049	0.05771
	Local Polynomial Regression	10313	0.00563	-0.00004	0.00005
	Horvitz Thompson Estimator	10313	0.05692	-0.02734	0.06427
MSE (koe)	Nadaraya Watson Estimator	10313	49,926.22	0.00003	34.38237
	Local Polynomial Regression	10313	0.104170373	0.00000	0.00002
	Horvitz Thompson Estimator	10313	56,455.50	0.17645	42.64341
Petroleum Demand (koe)	Nadaraya Watson Estimator	10313	3,061,485.90	165.01	427.56
	Local Polynomial Regression	10313	3,062,406.02	101.03	940.28
	Horvitz Thompson Estimator	10313	3,060,773.90	211.42	386.26

Table 4. *Final Petroleum Demand as estimated by the Local Polynomial Estimators for the Low, Middle and High Classes.*

Economic Class		Petroleum Demand (Sample)	Imputed petroleum demand	Total Estimate
Low	N	5	5,151	5,156
	Sum	1,09122	1,278,12486	1,279216.08
Middle	N	3	2,893	2,896
	Sum	1,00339	941,574.84	942578.23
Low	N	2	2269	2271
	Sum	994.49	842706.32	843700.81
Middle	N	10	10,313	10,323
	Sum	3.089.10	3.062.406.02	3.065.495.12

5. Conclusion

From this study, Local Polynomial Estimator performed better with less bias at the boundary as compared to the Nadaraya-Watson and the Horvitz Thompson Estimators. This study found that the bulk of petroleum demand in the county (49.85%) was consumed by the low class while the high class consumed less (22.09%). The remaining demand of 28.06% was attributed to the middle class.

The bias of local polynomial estimation was between -0.00004 to 0.00005 koe with the bulk of its bias lying around the zero mark. The boundary bias captured all the ten clusters unlike the Nadaraya-Watson estimator. The Local polynomial estimator was consistent at the boundary with an estimated total demand of 3,061,485.90koe.

Acknowledgements

The authors wish to acknowledge the Kenya National Bureau of Statistics for the grant provided towards this study.

References

[1] Christian Fürber (2016) Data Quality Management with Semantic Technologies Springer Gabler.

[2] Roderick J. A, Donald B. Rubin (2002) Statistical Analysis with Missing Data, Wiley-Interscience.

[3] Wanishsakpong, W., & Notodiputro, K. A. (2018). Locally weighted scatter ‐ plot smoothing for analysing temperature changes and patterns in A ustralia. *Meteorological Applications*, *25* (3), 357-364.

[4] Jack E. Olson (2003) Data Quality: The Accuracy Dimension (The Morgan Kaufmann Series in Data Management Systems) 1st Edition Morgan Kaufmann.

[5] Alexandra, A., Megan, D., Elizabeth, D. and Shivani, M. (2015). City-Level Energy Decision Making: Data Use in Energy Planning, Implementation, and Evaluation in U.S. Cities NREL is a national laboratory of the U.S. Department of Energy Office of Energy Efficiency & Renewable EnergyOperated by the Alliance for Sustainable Energy, LLCT report.

[6] Kihara, P. N. (2013). Estimation of Finite Population Total in the Face of Missing Values Using Model Calibration and Model Assistance on Semiparametric and Nonparametric Models. PhD thesis.

[7] Rüeger, S., McDaid, A., & Kutalik, Z. (2018). Improved imputation of summary statistics for admixed populations. *bioRxiv*, 203927.

[8] Mbugua, L. (2014). Modeling energy demand using nonparametric and extreme value theory. Lambert Academic Publishing.

[9] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics ISBN: 9780470316696 |DOI:10.1002/9780470316696.

[10] Schafer, J. L. (1999) Multiple Imputation: A Primer. Statistical MethodsinMedicalResearch, 8, 3-15. http://dx.doi.org/10.1191/096228099671525676.

[11] Schafer, J. L and John W. G. (2002) Missing Data: Our View of the State of the Art. Psychological Methods. The American Psychological Association, 7 (2), 147–177

[12] Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87, 998-1004.

[13] Fan, J and Gijbels, I (2003). Local polynomial modeling and its application. Chapman and Hall.

[14] Ruppert, D and Wand, M. P (1994). Multivariate weighted least squares regression. Ann. Statist. 22, 1346–70.

[15] Horvitz, D., and Thompson, D. (1952) A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association,* 47:663-685.

[16] Breidt, F. J., Opsomer, J. D., Johnson, A. A., and Ranalli, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. Survey Methodology, 33 (1), 35.

[17] Cochran, W. G. (1977). Sampling techniques-3. New York, NY (USA) Wiley.

[18] Pyeye, S. (2018). Imputation Based On Local Polynomial Regression for Nonmonotone Nonrespondents in Longitudinal Surveys (Doctoral dissertation, JKUAT-PAUSTI).

[19] Fritz, M. (2019). Steady state adjusting trends using a data-driven local polynomial regression. *Economic Modelling*.

[20] Cattaneo, M. D., Jansson, M., & Ma, X. (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association*, (just-accepted), 1-11.