Nonparametric Model-Based Predictive Estimation in Survey Sampling

April Anne H. Kwong-Aquino
University of the Philippines Visayas

A nonparametric model-based estimator of the population total is proposed. The sample data along with the auxiliary information are used in fitting a generalized additive model that is then used in reconstructing the unknown population. The estimates of the population parameters are computed from the predicted population values (for the unsampled part of the population) and the sample values. A simulation study designed to account for different association patterns between the target variable and the auxiliary variable, population size, and sample size was conducted to evaluate the proposed procedure. The method is robust to data-generating model form, population size, and sampling rate, and is generally superior to design-unbiased estimators.

Keywords: model-based estimation, predictive estimation, nonparametric regression, additive model

1. Introduction

In survey data analysis, one of the primary objectives is the estimation of population characteristics and often, design-unbiased, model-assisted, or model-based techniques are used. Design-unbiased methods of estimation are dependent on the sampling distribution induced by the sample selection process where knowledge of the population structure is very important. Design-based estimation requires availability of population frame and sampling weights (alternatively, selection probabilities). Oftentimes, complete and reliable information about the population needed in frame construction is difficult to secure. One way to address this problem is to use model-based estimation techniques. Model-based estimation procedure does not completely depend on the population frame and does not require knowledge of weights to account for the unsampled segment of the population. Instead, the auxiliary information is used to predict the unsampled values. This

method requires the estimation of the relationship between the two variables by fitting a regression equation using the pairs (x_i, y_i), for $i \in S$, (Barrios, 2007). Model-assisted approach integrates the design-unbiased and model-based approaches. Estimation does not rely on model assumptions and inferences are based on the survey design alone, however, models are used to specify the parameters of interest (Lohr, 1999). (Rueda and Sanchez-Borrego, 2009) noted the advantage of model-based estimation when there is a strong linear relationship between the target and auxiliary variables.

Model-based inferences have been evaluated using both parametric and nonparametric models. Parametric models necessitate that model assumptions are met for the inferences to be valid. Under this approach, the accuracy of the specified model is a major concern. Also, conditions like linearity, normality and independence of the error terms must be satisfied. However, deviations from these assumptions are to be expected. These deviations and misspecifications lead to erroneous inferences. Robustness of estimation procedures to model assumptions is, thus, desirable. The use of nonparametric methods can possibly achieve this as they allow the data to dictate the relationship curve of the variables (Eubank, 1998). This is an appropriate alternative when there is little a priori information on the structure of the relationship or even when there is doubt about the validity of the parametric model.

Barrios (2007) proposed a model-based estimation technique in estimating the population total for variables which are linearly related to an auxiliary variable. The estimator is given by $\hat{T} = \sum_{i \in S} y_i + \sum_{j \notin S} \hat{y}_j$ where y_i are the sample values and $\hat{y}_j$ are the predicted values of the unsampled population units. Simulation scenarios that included symmetric and skewed populations were conducted. The normal regression model was used for the symmetric population while poisson regression with log link function was used for the skewed population. The proposed estimation procedure yielded comparable results to design-based estimation for small and large populations.

An estimator of the population mean was proposed by Rueda and Sanchez-Borrego (2009) using local polynomial regression and was compared to several existing methods of estimation using simulated populations. The estimator of the population mean is given by $\bar{y}_{MB} = f\bar{y}_S + (1-f)\frac{1}{N-n}\sum_{j \in \bar{S}} \hat{m}_j$ where $\bar{y}_S = \frac{1}{n}\sum_{i \in S} y_i$ is the sample mean and $\hat{m}_j$ are the predicted values of the unsampled part of the population. The $\hat{m}_j$ are computed using a local polynomial regression model that was fitted to the sample data on the pairs (x_i, y_i). The proposed method exhibited satisfactory performance relative to design-based estimators as the sample size decreases.

The combination of the concept of model-based estimation and nonparametric methods promises several advantages. Sampling and estimation can be done even without a reliable and complete population frame. Prior information on the attributes of the population is also not a primary concern.

2. Estimation of the Population Total

Assume that the response variable *Y* is related to an auxiliary variable whose census is known. Assume further that the values of *Y* corresponding to the maximum and minimum values of *X* are known or can at least be estimated prior to analysis. The variable of interest (*Y*) and the auxiliary variable (*X*) are assumed to be related through the function $y = f(x) + \varepsilon$. The following algorithm is proposed to estimate the population total of *Y*.

1. Given that a census on X is available, obtain sample information on *Y* and X (S).
2. Estimate $f(x)$ through a generalized additive model to the pairs (x_i, y_i), $i \in S$.
3. Predict the unsampled part of the population using the estimated generalized additive model, i.e., $\hat{y}_j = \hat{f}(x)$ for $j \notin S$.
4. Combine the sample values y_i, $i \in S$ and predicted values $\hat{y}_j$, to recreate the population.
5. Estimate the population total using $\hat{T} = \sum_{i \in S} y_i + \sum_{j \notin S} \hat{y}_j$.

The estimator is called Nonparametric Model-Based Estimator (NMBE).

3. Estimation of the Standard Error of $\hat{T}$

The standard error of the proposed estimator $\hat{T} = \sum_{i \in S} y_i + \sum_{j \notin S} \hat{y}_j$ can be computed using the bootstrap. Two hundred replicates are selected from the recreated population given a sampling rate of 50%. The statistic $\tilde{T}_i = N\overline{Y}_i$ for every $i = 1, 2, \ldots, 200$ is computed along with $\hat{\sigma}(\tilde{T}) = \sqrt{\frac{\sum_{i=1}^{n}\left(\tilde{T}_i - \bar{\tilde{T}}\right)^2}{n-1}}$ where $\bar{\tilde{T}}$ is the mean of the $\tilde{T}_i$'s and n is the number of replicates (200). The computed value of $\hat{\sigma}(\tilde{T})$ estimates the standard error of $\hat{T}$. This is compared to the estimate of the standard error of the design-unbiased estimate of the population total, $\hat{\sigma}_{SRS} = \sqrt{\left(\frac{N-n}{N-1}\right)}\frac{Ns}{\sqrt{n}}$.

4. Simulation Study

The algorithm above is implemented on several simulation settings to evaluate the performance of the method. Each scenario postulates a model $y = f(x) + k * \varepsilon$, where *f(x)* is either linear or nonlinear in *x*. Populations of variables exhibiting linear, quadratic, and exponential relationships are simulated. The equations used are:

Linear	$y = 1.35x + k * e$
Quadratic	$y = -0.1x^2 + 10x + k * e$
Exponential	$y = 10\exp(x/25) + k * e$

These relationships represent some possible association patterns between variables. The constant multiplier (k) to the error term is used to induce varying degrees of model misspecification, small k implies no or minimal misspecification, while large k implies severe misspecification. For each model above, population size, sampling rate, and multiplier on the error term were varied. Small and large finite populations are represented by populations of sizes $N = 1000$ and $N = 10000$, respectively, from where random samples are taken given the sampling rates: 1%, 3%, 5%, 10%, and 20%. Error terms are assumed to follow the standard normal distribution with multipliers (k) set to 1, 5, 8.75 and 10. For $k > 1$, the error is bloated, destroying the fit of *f(x)* to the data. This introduces additional variability in *Y* that is explained by components other than *f(x)* and lessens the capacity of *X* to predict the values of *Y*. Additional considerations include setting the coefficient of *X* in the linear equation to 1.35 in order to generate data values with $r \approx 0.95$ for $k = 1$, $r \approx 0.50$ for $k = 5$ and $r \approx 0.30$ for $k = 8.75$ when variance of *X* is set at 5. These values of the correlation coefficient correspond to strong, average, and weak linear relationships, respectively. The auxiliary variable, *X*, was simulated to follow the normal distribution with mean equal to 50. To investigate the effect of the variance of *X* on the efficiency of the estimates, variances are set to 5, 25 and 225, and 400.

The frame problem posed by some population units not being accessible to sampling is also considered. For each simulation scenario presented above, NMBE estimates using samples from the middle 50% of the population are also calculated. Under this case, the unsampled segment of the population includes the lower and upper 25% as well as the units in the middle 50% that were not captured during sampling.

To evaluate the performance of the proposed estimator, the absolute percent difference is computed from $PD_{est} = \left(abs\left(\frac{\hat{T}_{est} - T}{T} \right) \right) * 100\%$, where T is the true population total. The percentage advantage of NMBE estimates over design-unbiased estimates is further computed from $PA = \frac{\overline{P}D_{SRS} - \overline{P}D_{NMBE}}{\overline{P}D_{SRS}} * 100\%$, where $\overline{P}D_{NMBE}$

is the average absolute percent difference between the NMBE estimate and the true population total and $\overline{PD}_{SRS}$ is the average absolute percent difference between the SRS estimate and the true population total.

5. Results and Discussions

The different scenarios in the simulation study are considered in order to vary the amount of variability in the population as measured by the coefficient of variation. This characteristic of the population has an effect on the performance and even on the choice of estimation procedures. High coefficients of variation indicate high heterogeneity of population values. For such cases, estimation can be more complex and costly as a large sample is needed to ensure that the variability pattern is captured adequately by the sample. On the contrary, a small sample would suffice to represent a population with low coefficients of variation.

We summarized in Tables 1-3 some values of population parameters resulting from the restrictions imposed during the simulation for $N = 1000$. Similar values are generated for $N = 10000$.

Table 1 Coefficient of Variation (CV) of $Y = 1.35X + k*e$ and Correlation Coefficient Across Varying Values of *Var*(*X*) and *k*

Var(X)	5			25			225			400		
k	1	5	8.75	1	5	8.75	1	5	8.75	1	5	8.75
r	0.95	0.50	0.30	0.99	0.80	0.60	0.99	0.97	0.92	0.99	0.98	0.95
CV of *Y*	4.63	8.34	13.24	10.01	12.09	15.81	29.95	30.59	32.14	40.01	40.44	41.58

Table 2 Coefficient of Variation (CV) of $Y = -0.1X^2 + 10X + k*e$ Across Varying Values of *Var(X)* and *k*

Var(X)	5			25			225			400		
k	1	5	10	1	5	10	1	5	10	1	5	10
CV of *Y*	0.48	2.04	4.06	1.46	2.45	4.28	13.81	13.91	14.36	26.61	26.64	26.86

Table 3 Coefficient of Variation (CV) of $Y = 10exp(x/25) + k*e$ Across Varying Values of *Var(X)* and *k*

Var(X)	5			25			225			400		
k	1	5	10	1	5	10	1	5	10	1	5	10
CV of *Y*	9.00	211.45	16.61	19.92	21.34	24.66	62.62	63.31	64.63	87.33	87.90	88.85

5.1 Effect of model form

We summarized in Table 4 the average absolute percent differences of both the design-unbiased (SRS) and NMBE estimates from the true population total over different model forms. The percentage advantage (*PA*) of NMBE over SRS is also summarized in Table 5. This is evaluated only for NMBE estimates that utilized the entire population.

The average absolute percent differences of NMBE estimates from the true population total are lower than that of the design-unbiased estimates for all model forms and across sampling rates. Moreover, the contrast becomes clearer as the sampling rate decreases. With the linear association, the NMBE estimates are better for sampling rates up to 5% and comparable to the design-unbiased estimate for much higher sampling rates. The same trend can be seen for the quadratic model where the NMBE is better for sampling rate 1% only and exhibited performance similar to the design-unbiased estimator for sampling rates 3% up to 20%. For the exponential function, the proposed estimator is superior for all sampling rates.

The dissimilarity in the percent differences of the estimates is larger for the linear and exponential model forms with NMBE estimates exhibiting advantage over design-unbiased estimates. These patterns can be attributed more to the coefficients of variation of the population than to the model form. For the linear and the exponential case, coefficients of variation range from 5% to 89%. On the other hand, resulting coefficients of variation for the quadratic relationship are from 0.5% to 27% only. Coefficients of variation indicate the level of homogeneity of the population values. SRS is more appropriate for homogeneous populations, that is, populations with low coefficients of variation. Hence, they are expected to perform well in such scenarios. The design, however, may fail to give accurate estimates for heterogeneous populations or populations with high coefficients of variation. By inspection of the values, the NMBE estimates seem to be unaffected by this population characteristic.

The model forms considered in this paper yield different types of population characteristic in relation to symmetry. The linear form outputs symmetric data while the quadratic form and exponential form introduce negative skewness and positive skewness, respectively. The percentage advantage of NMBE over the SRS-unbiased estimator does not change much with respect to model form. This indicates that the model forms, and as a consequence type of skewness, does not necessarily determine the cases for which NMBE estimates are more superior.

As the sample size increases, the accuracy of SRS estimates also increases. The NMBE estimates, however, are relatively robust to sample size changes. In addition, the average absolute percent differences of NMBE estimates are significantly lower than that of design-based estimates for smaller samples, highlighting the advantage

Table 4 Average Absolute Percent Difference of the Estimates With Respect to Model Form

	1%			3%			5%			10%			20%		
	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)
Linear	4.09	1.30	0.97	2.30	0.72	0.50	1.75	0.53	0.41	1.24	0.37	0.31	0.73	0.24	0.25
Quadratic	2.02	0.57	0.78	1.12	0.32	0.55	0.94	0.27	0.44	0.60	0.16	0.25	0.33	0.10	0.22
Exponential	8.14	1.67	5.69	4.52	0.97	5.16	3.30	0.69	4.81	2.29	0.48	4.32	1.49	0.30	4.36

Table 5 Percentage Advantage of NMBE Estimate over SRS-unbiased Estimate With Respect to Model Form

	1%	3%	5%	10%	20%
Linear	68.192	68.823	69.903	70.129	67.123
Quadratic	71.726	71.505	71.868	72.833	69.970
Exponential	79.496	78.526	78.963	79.214	79.584

of NMBE in small samples as also pointed out in (Rueda and Sanchez-Borrego, 2009).

For the Linear model, it can be observed that the percent differences of the NMBE estimates from the true population total when sampling is only from the middle 50% of the population is lower when compared to the percent differences of the other two estimates (SRS and NMBE) which are based on the entire population. The auxiliary variable *X* was simulated to follow the normal distribution. Thus, the Linear model $y = 1.35x + k * e$ also yielded a symmetric distribution for Y. Due to the symmetry of *Y*, sampling on the middle 50% decreased the percent difference of the estimate. When only the middle 50% of the population is sampled, the likelihood that values in the neighborhood of the mean are selected increases. As a result, the estimates are also more accurate.

The quadratic model ($y = -0.1x^2 + 10x + k * e$) and exponential model ($y = 10\exp(x/25) + k * e$) resulted to a skewed distribution for *Y*. Sampling on the middle 50% yields less accurate estimates as extreme values that caused the skewness in *Y* did not have representation in the sample of paired observations (x_i, y_i) .

5.2 Effect of variance of X

We summarized in Table 6 the average absolute percent differences of both design-unbiased (SRS) and NMBE estimates from the true population total over varying variance of *X*. The percentage advantage of NMBE (considering the entire population) over SRS are presented in Table 7.

From Table 6, there are slight changes in the NMBE estimates across different variances of *X*. However, estimates of SRS are highly variable with respect to these adjustments. There is a direct proportional relationship between the variances of the auxiliary and the target variable, an increase in the variation of *X* adds to the heterogeneity of the population of *Y* and as a consequence, affects the performance of SRS estimates. As it increases further, NMBE estimates are more accurate than SRS estimates even as the sampling rate becomes much larger.

The percentage advantage of NMBE does not change much with respect to sampling rate. However, there is an apparent increase in that measure as more variability is introduced to the auxiliary variable. By increasing the variance in *X* while holding all other simulation parameters constant, proportion of variability in *f(X)* that accounts for the variability in *Y* increases. This is tantamount to increasing the predictive ability of the auxiliary variable thus resulting to an increase in the advantage of the proposed estimator.

For low variations in *X*, the simulated frame problem (only the middle 50% of the population is accessible to sampling) actually yields better NMBE estimates. The variability in *Y* proportionally changes with respect to the variability in *X*. When

Table 6 Average Absolute Percent Difference of the Estimates With Respect to Variance of *X*

	1%			3%			5%			10%			20%		
	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)
5	1.370	0.936	0.510	0.749	0.53	0.272	0.55	0.388	0.208	0.394	0.279	0.141	0.255	0.186	0.092
25	2.222	0.929	0.714	1.198	0.529	0.411	0.903	0.387	0.302	0.633	0.277	0.227	0.397	0.186	0.155
225	6.349	1.141	2.800	3.543	0.654	2.388	2.683	0.487	2.157	1.843	0.329	1.826	1.146	0.224	1.734
400	9.042	1.712	5.886	5.091	0.961	5.220	3.855	0.719	4.882	2.667	0.461	4.309	1.651	0.262	4.451

Table 7 Percentage Advantage of NMBE Estimate over SRS-unbiased Estimate With Respect to Var(*X*)

	1%	3%5%	10%	20%	
5	31.679	29.239	29.455	29.188	27.059
25	58.191	55.843	57.143	56.240	53.149
225	82.029	81.541	81.849	82.149	80.45
400	81.066	81.124	81.349	82.715	84.131

there is little variation in X and when the Y values also do not vary much, sampling from only the middle 50% is sufficient to acquire a representative of the population. Given a small variance of X and Y, this scheme is actually advantageous over sampling from the entire population. There is higher probability of sampling around the mean when we focus sampling on the middle 50% than when the entire population is considered. With the latter case, the extreme cases (which are uncommon when the variation is low) still have a nonzero probability of inclusion. When captured during sampling, these extreme values can pull the sample mean in either direction depending on the direction of the extreme case, if it is significantly higher or lower than the mean, resulting to bias in the estimation. Sampling from the middle 50% removes this possible source of inaccuracy. However, the same sampling restriction will not yield desirable results if the population variability is high as can be seen in Table 6 for variances of X equal to 225 and 400. The percent difference of the estimates from the true population total significantly increased for higher levels of heterogeneity in X and, as a consequence, in Y. Since the spread of the population values is wide, sampling from only the middle 50% will not be enough to get a good representative of the population. For the symmetric case with high variability, density of the values will not vary much as you move in either direction away from the mean. Sampling from the middle 50% disregards this and assigns zero inclusion probability to values which are farther from the mean. This will result to lesser accuracy of the estimates.

5.3 Effect of population size

Table 8 shows the average absolute percent differences of both design-unbiased (SRS) and NMBE estimates from the true population total with respect to population size. The percentage advantage of NMBE estimates (sampling over the entire population) over SRS estimates are shown in Table 9.

Simulated populations that vary with respect to population size alone have the same amount of variation. Hence, the same sample size irrespective of the population size will result to similar inferences. This can be observed for the estimates given the sampling rate of 10% for population size of 1000 and for the estimates given the sampling rate 1% for population size of 10000. The absolute percent differences of SRS estimates are similar (2.089 for $N = 1000$ and 2.263 for $N = 10000$). The same is true for the NMBE estimates (0.509 for $N = 1000$ and 0.630 for $N = 10000$).

Regardless of the population size, NMBE is more advantageous than SRS (Table 9). Moreover, there are no notable differences in the percentage advantage of NMBE over the SRS-unbiased estimator considering changes in the population size. This indicates that population size does not contribute to the efficiency of the NMBE estimate relative to the efficiency of the SRS-unbiased estimate.

Even though sampling is only from the middle 50%, NMBE still performed better relative to SRS-unbiased estimation (based on the entire population) for small

Table 8 Average Absolute Percent Difference of the Estimates With Respect to Population Size

	1%			3%			5%			10%			20%		
	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)
1000	7.13	1.729	2.401	4.054	0.991	1.939	2.967	0.743	1.778	2.089	0.509	1.621	1.275	0.317	1.404
10000	2.362	0.630	2.555	1.236	0.347	2.207	1.028	0.248	1.996	0.68	0.164	1.630	0.45	0.112	1.812

Table 9 Percentage Advantage of NMBE Estimate over SRS-unbiased Estimate With Respect to Population Size

	1%	3%	5%	10%	20%
1000	75.750	75.555	74.958	75.634	75.137
10000	73.328	71.926	75.875	75.882	75.111

samples from the smaller population. For the larger population, the performance of the NMBE estimates and that of the SRS-based estimates are comparable when sampling rate is at least 10%. However, for other sampling rates, the design-unbiased estimates have lower percent differences.

5.4 Effect of model fit

Presented in Table 10 are the average absolute percent differences of both design-unbiased (SRS) and NMBE estimates from the population total with respect to the model fit. The percentage advantage of NMBE estimates (using the entire population) over SRS estimates are shown in Table 11.

There are slight changes in the average absolute percent difference of both SRS and NMBE estimates with respect to the increases in the error multiplier *k*. Increasing *k* decreases the amount of variation in *Y* that is explained by *f(X)*. The NMBE estimates are model dependent, thus, their performance is affected by changes in the model fit. For $k = 1$, NMBE estimates yield better performance for sampling rates of at most 10% while for larger values of *k* (greater than 5), NBME estimates are more accurate for sampling rates of at most 5%.

The percent difference of the NMBE estimates from the actual population total increases as the error multiplier also increases. These estimates are model-based and any changes in the capacity of the auxiliary variable *X* to predict values of *Y* will affect their accuracy. These changes are incorporated during simulation through the error multiplier *k*. Increasing k decreases the prediction capability of *X* and as a consequence, increases the percent difference of the estimate from the population value. It can be observed that the increments are larger when the sampling rate is small suggesting that increasing the sample size reduces the effect of the error multiplier.

There is no clear pattern in the percentage advantage of NMBE estimates with respect to the sampling rates. Model-dependent methods rely on the relationship between variable of interest and auxiliary variable to capture the trends in the data. They usually fail as estimation tools when model does not fit the data well (Kalton, 2002). This stresses the importance of that assumption even when using nonparametric methods. While they might be robust to model assumptions, the minimum requirement of an association between the variables *X* and *Y* still needs to be met. Otherwise, the performance of nonparametric model-based techniques will not be optimal.

For sampling from only the middle 50%, NMBE still performed better across all the given values of the error multiplier for sampling rate of 1%. The completely random nature of SRS sampling translates to better representation of the population when the samples are large. Thus for small sample cases, the performance of NMBE

Table 10 Average Absolute Percent Difference of the Estimates With Respect to the Value of the Error Multiplier, *k*

	1%			3%			5%			10%			20%		
	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)	SRS	NMBE	NMBE (50%)
1	4.46	0.572	1.887	2.516	0.326	1.682	1.895	0.242	1.552	1.311	0.156	1.339	0.810	0.087	1.847
5	4.697	1.142	2.281	2.611	0.646	1.999	1.99	0.480	1.846	1.367	0.324	1.631	0.851	0.206	1.358
>8.75	5.081	1.825	2.877	2.808	1.035	2.293	2.123	0.764	2.041	1.476	0.530	1.714	0.927	0.350	1.428

Table 11 Percentage Advantage of NMBE Estimate over SRS-unbiased Estimate With Respect to the Value of the Error Multiplier, *k*

	1%	3%	5%	10%	20%
1	87.175	87.043	87.230	77.101	89.259
5	75.687	75.259	75.696	76.298	75.793
>8.75	64.082	63.141	64.013	64.092	62.244

relies heavily on the relationship between *X* and *Y*. For all the other scenarios, NMBE estimates using only the middle 50% of the population are comparable to SRS-estimates based on the entire population.

5.5 Standard error of NMBE estimates

For each simulation scenario, the estimated standard error and coefficient of variation of the NMBE estimate is lower than that of the SRS estimate. According to Lohr (1999), the standard errors of estimates under model based approaches are generally lower than those of design based estimates. This can be attributed to the differences in the way the standard errors are computed. The variance of model-based estimates is the average squared deviation of the estimate from its expected value with the average computed over all possible values that the model can generate. While for the design-based estimates, the average is over all possible samples using the specified design.

The estimate of the standard error and, correspondingly, the coefficient of variation (CV) of the design-unbiased estimate is a function of the sample variance and the sampling rate. The decreasing trend in the CV as the sample size increases can be observed in the results. However, the effect of the sample size on the CV of the proposed estimator is negligible. The coefficients of variation of the latter estimator also exhibited robustness to the variation in the auxiliary variable and model fit.

6. Conclusions

The performance of the nonparametric model-based estimates (NMBE) is at the least comparable to SRS estimates. The NMBE estimates are superior to design-unbiased (SRS) estimates when the population is very heterogeneous and when a high proportion of variation in *Y* is accounted for by variation in *f(X)*, i.e., when there is a relationship between Y and X. NMBE also provide an alternative strategy in estimating population total where sampling suffers from severe frame problem.

REFERENCES

BARRIOS, E., 2007, Model Based Predictive Estimation with Coverage Error, Bulletin of the 57th Session of the International Statistical Institute, Portugal.

RUEDA, M. and SANCHEZ-BORREGO I.R., 2009, A Predictive Estimator of Finite Population Mean using Nonparametric Regression, *Comput Stat*, 24:1-14.

KALTON, G., 2002, Models in the Practice of Survey Sampling (Revisited), *Journal of Official Statistics* 18:129-154.

LOHR, S., 1999, *Sampling: Design and Analysis*. Brooks/Cole.

EUBANK, R., 1998, *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, Inc.