SCHOOL OF STATISTICS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

Nonparametric Model-Based Estimation In Data Mining

by

April Anne H. Kwong[1]
and
Erniel B. Barrios[2]

UPSS Working Paper No. 2010-13
September 2010

School of Statistics
Ramon Magsaysay Avenue
U.P. Diliman, Quezon City
Telefax: 928-08-81
Email: updstat@yahoo.com

[1] Instructor, University of the Philippines in the Visayas

2 Professor, School of Statistics, University of the Philippines Diliman

Abstract

Probability sampling in finite populations are completely dependent on the availability of a reliable frame. In market research, especially for new products/services, the frame that enumerates the target market is not available. Official statistics like census and survey data are regularly collected by the Philippine Statistical System. The public use files of these data systems can be potentially beneficial among researches in the business sector. Using the weighted household-level data in the 2003 Family Income and Expenditure Survey, The proposed nonparametric model-based estimation procedure is used to estimate the market size for food items and some of its components. Model-based estimation is viewed in the context of re-sampling methods to estimate the population total. Even if the sample is drawn only from a small part of the population, model-based estimates are superior or at least comparable to design-based estimates especially for small populations. In symmetric populations, the choice of an auxiliary variable (predictor) is important but in a skewed population, performance of model-based estimator is robust to the relationship between the target variable and the auxiliary predictor. The bootstrap sampling errors are generally lower than the design-unbiased sampling errors.

Keywords: Model-Based Estimation, Generalized Additive Model, Nonparametric Regression

1. Introduction

In survey sampling, one major goal is the estimation of estimation of population characteristics like the mean and total. There are many estimation procedures available and the more popular are design-unbiased, model-assisted, and model-based techniques. Design-unbiased methods of estimation are dependent on the sampling distribution induced by the sample selection process and access on the population frame is crucial under this paradigm. Design-based estimation relies on the population frame and sampling weights. In practice, complete and reliable information about the population is difficult or expensive to access. Thus, to avoid the difficulty of coming up with the stringent requirements of design-based estimation, model-based estimation techniques can be used instead. Model-based estimation procedure does not completely depend on the population frame and does not use weights to account for the unsampled segment of the population. The auxiliary information is used to predict the unsampled values. The process involves estimation of the relationship between the two variables by fitting a regression model using the pairs (x_i, y_i), $i \in S$ and predicting the values of the unsampled part to complete the information that will be used in characterizing the population, see for example, (Barrios, 2007). The model-assisted approach integrates the design-unbiased and model-based methods. 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). In a study on model-based estimation (Rueda and Sanchez-Borrego, 2009), the population in the year 1970 of 304 counties in North and South Carolina and Georgia was estimated using the number of households in 1960 in the same location as auxiliary variable. The model-based

technique performed well for this case due to the strong linear relationship between the target and auxiliary variables.

Model-based inferences have used both parametric and nonparametric models. Parametric models necessitate that certain assumptions are met for the inference to be valid. Under this approach, the accuracy of the specified model is an important consideration, e.g, normality and independence of the error terms. In practice, deviations from these assumptions are very common. These deviations and model misspecifications lead to erroneous inferences. Thus, robustness of estimation procedures to model assumptions is desirable. The use of nonparametric methods can possibly mediate in this case since it allows the data to dictate the form of the relationship among 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 for 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 the y_i are the sample values and $\hat{y}_j$ are the predicted values of the unsampled population units. The method was evaluated using intensive simulation scenarios including symmetric and skewed populations. Modelling for the symmetric population used normal regression model, while poisson regression with log link function was used for the skewed population. The method performed better than design-based estimation techniques for small and large populations.

An estimator of the population mean was proposed by. (Rueda and Sanchez-Borrego, 2009) proposed an estimator of the population mean based on the 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 $\bar{y}_S = \frac{1}{n}\sum_{i \in S} y_i$ and $\hat{m}_j$ are the predicted values of the unsampled part of the population. The $\hat{m}_j$'s 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 was found to exhibit satisfactory performance relative to design-based estimators as the sample size decreases.

Combining the concept of model-based estimation and nonparametric methods, similar to the work of (Rueda and Sanchez-Borrego, 2009), 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. This study is motivated by these advantages and examines how other nonparametric methods of regression can be invoked as tools in survey data analysis.

2. The Generalized Additive Model

An additive model generalizes the multiple regression model. It maintains the additive nature but replaces the terms of the linear equation with smooth functions of the independent variable instead of fixed parametric functions. The regression model is given by

$$E(Y \mid X_1,..., X_p) = f(X_1,..., X_p) = s_0 + s_1(X_1) + ... + s_p(X_p).$$

Where the smooth functions, s_i, are estimated using nonparametric methods.

Generalized linear models (GLM) assume that the dependent variable is related to a linear combination of the independent variable though a link function. The conditional mean is modelled as $g(E(Y \mid X_1,..., X_p)) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$. This form allows the response variable to assume distributions other than the normal distribution (StatSoft, Inc., 2010).

A generalized additive model (GAM) integrates the additive model and the generalized linear model. GAM links the mean of the response variable to an additive predictor through the function, $g(E(Y \mid X_1,..., X_p)) = s_0 + \sum_{i=1}^{p} s_i(X_i)$. The link function connects the additive and the random components of the model. The additive component is the quantity $\eta = s_0 + \sum_{i=1}^{p} s_i(X_i)$ where the terms are smooth functions of the independent variable and are also estimated in a nonparametric manner as in the case of additive models. The random component is the response variable *Y* with a probability distribution that can be written in the form $f_Y(y;\theta;\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$ where θ and ϕ are the location and scale parameters, respectively. That is, the random variable *Y* can have any distribution that belongs to the exponential family (Hastie and Tibshirani, 1990).

Generalized additive models allow the mean of the response variable to depend on an additive predictor through the link function, $g(E(Y \mid X_1,..., X_p)) = s_0 + \sum_{i=1}^{p} s_i(X_i)$. This translates to allowing the distribution of the dependent variable to be any distribution which belongs to the exponential family of distributions. Furthermore, relationships and structures among variables are not restricted to linearity as the smooth functions s_0, s_1, ..., s_p are not prescribed any form before model fitting and are estimated in a nonparametric manner.

3. Estimation of the Population Total

The method proposed in this paper assumes that the response variable *Y* is related in some fashion to an auxiliary variable whose census is known. Furthermore, the values of *Y* corresponding to the maximum and minimum values of *X* must be known or can at least be estimated prior to analysis. The data consisting of the random sample and the values of *Y* at the extremes of *Y* is called augmented sample.

We propose that a generalized additive model (GAM) is used to fit the relationship between x and y. The fitted model on the relationship between *x* and *y* is then used to predict the values of the unsampled part of the population.

In survey estimation, the primary objective is to make generalizations about the population using sample data. The population mean, total, and proportion are some parameters that researchers estimate in order to develop policies or identify markets for certain products. Sampling basically divides the population into the sampled and the unsampled segments. Design-based estimation uses only sample information and employs weights to account for the unsampled segment. On the other hand, the model-based approach predicts the values of the unsampled part using an estimate of the relationship between the variable of interest and an auxiliary variable. The population is recreated by combining the sample and the predicted values of the unsampled part. Barrios (2007) and Rueda and Sanchez – Borrego (2009) have investigated performances of this technique using parametric and nonparametric modelling of the association between variables.

The variable of interest (*Y*) and the auxiliary variable (*X*) are assumed to be related through the function $y = f(x) + \varepsilon$. An estimation procedure called Nonparametric Model-Based Estimator (NMBE) is proposed following the algorithm below as a procedure for estimating the population total *Y*.

1. Obtain sample information on *Y* and a census on *X*.
2. Estimate *f* (*x*) by fitting a generalized additive model to the pairs (x_i, y_i), $i \in S$.
3. Predict the unsampled part of the population using the fitted generalized additive model. That is, compute $\hat{y}_j = \hat{f}(x)$ where $j \notin S$.
4. Combine the sample values y_i, $i \in S$ and predicted values $\hat{y}_j$, $j \notin S$ 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 standard error of the proposed estimator $\hat{T} = \sum_{i \in S} y_i + \sum_{j \notin S} \hat{y}_j$ is done using the bootstrap. Two hundred replicates are selected from the recreated population given a sampling rate of 50%. The statistic $\tilde{T}_i = N\overline{X}_i$ for every *i* = 1, 2, ... , 200 is computed along with $\hat{\sigma}(\tilde{T}) = \sqrt{\frac{\sum_{i=1}^{B} \left(\tilde{T}_i - \bar{\tilde{T}}\right)^2}{B-1}}$ where $\bar{\tilde{T}}$ is the mean of the $\tilde{T}_i$'s and *B* is the number of replicates, 200. The computed value of $\hat{\sigma}(\tilde{T})$ estimates the standard error of $\hat{T}$.

4. Simulation Study

To evaluate the proposed estimation procedure, a simulation study is conducted. 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 following 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. For each of these equations, the following quantities are made to vary: population size, sampling rate, and multiplier (k) on the error term. Small and large finite populations are represented by populations of sizes *N* = 1000 and *N* = 10000, respectively, from where random samples are drawn given the sampling rates: 1%, 3%, 5%, 10%, and 20%. Error terms are generated from the standard normal distribution with multipliers set to 1, 5, 8.75 and 10. For *k* > 1, the error is magnified, destroying the fit of $f(x)$ to the data. This introduces additional variability in *Y* that is not explained by components $f(x)$ and lessens the capacity of *X* to predict the values of *Y*. Additional scenarios 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 $k = 8.75$. 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 50 and variances at 5, 25, 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 nonsampled 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 as defined by $PD_{est} = \left(abs\left(\frac{\hat{T}_{est} - T}{T} \right) \right) * 100\%$ where *T* is the true population total is computed for both the proposed estimator ($\hat{T}_{NMBE}$) and the design-unbiased estimator ($\hat{T}_{SRS}$). The estimator with the lesser percent difference is considered to have exhibited superior performance. The percentage advantage of NMBE estimates over design-unbiased estimates is $PA = \frac{\overline{P}D_{SRS} - \overline{P}D_{NMBE}}{\overline{P}D_{SRS}} * 100\%$,

where $\overline{P}D_{NMBE}$ = average absolute percent difference between the NMBE estimate and

the true population total and

$\overline{PD}_{SRS}$ = average absolute percent difference between the SRS estimate and the

true population total

are also assessed.

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 degree of 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 coefficient of variation.

Listed in the tables below are some values of population parameters resulting from the restrictions imposed during simulation for N = 1000. Similar values are generated for N = 10000.

Table 1. Coefficient of Variation and Pearson Correlation Coefficient of $Y = 1.35X + k*e$

Variance of X	5			25			225			400		
k	1	5	8.75	1	5	8.75	1	5	8.75	1	5	8.75
r	0.950	0.503	0.302	0.990	0.802	0.601	0.999	0.971	0.919	0.999	0.984	0.952
CV of *Y*	4.629	8.341	13.241	10.009	12.089	15.808	29.955	30.589	32.137	40.005	40.437	41.582

Table 2. Coefficient of Variation of $Y = -0.1X^2 + 10X + k*e$

Variance of X	5			25			225			400		
k	1	5	10	1	5	10	1	5	10	1	5	10
CV of *Y*	0.485	2.040	4.057	1.456	2.446	4.281	13.811	13.915	14.358	26.614	26.638	26.863

Table 3. Coefficient of Variation of $Y = 10exp(x/25) + k*e$

Variance of X	5			25			225			400		
k	1	5	10	1	5	10	1	5	10	1	5	10
CV of *Y*	9.002	11.453	16.611	19.915	21.336	24.663	62.624	63.315	64.627	87.333	87.897	88.846

5.1 Effect of Model Form

We presented in Table 4 the average absolute percent differences of both design-unbiased (SRS) and NMBE estimates from the true population total by model form. 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.

Table 4. Average Absolute Percent Difference of the Estimates

Data Generating Model Form	Sampling Rate														
	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	4.087	1.300	0.971	2.303	0.718	0.503	1.751	0.527	0.413	1.242	0.371	0.307	0.730	0.240	0.252
Quadratic	2.016	0.570	0.777	1.116	0.318	0.553	0.942	0.265	0.437	0.600	0.163	0.253	0.333	0.100	0.215
Exponential	8.135	1.668	5.685	4.517	0.970	5.163	3.299	0.694	4.810	2.290	0.476	4.316	1.489	0.304	4.357

Table 5. Percentage Advantage of NMBE Estimate over Design-Unbiased Estimate

Data Generating Model Form	Sampling Rate				
	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

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. NMBE is becoming more advantageous of design-unbiased estimator for smaller sampling rates. For the simulated population with linear association between y and x, 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 1% sampling rate similar performance similar as the design-unbiased estimator for sampling rates 3% to 20%. For the population generated from the exponential function, the NMBE 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 having the lower value. These patterns can be attributed 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% (Table 4.1 and Table 4.3). On the other hand, coefficients of variation for the quadratic relationship are from 0.5% to 27% only (Table 4.2). 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. SRS however, may fail to give accurate estimates for heterogeneous populations or populations with high coefficients of variation. The NMBE estimates are relatively robust to population heterogeneity.

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 design-unbiased estimator does not change much with respect to model form.

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 of NMBE in small samples similar to the results of (Rueda and Sanchez – Borrego, 2009).

For the population generated from a 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 yield a symmetric distribution for Y. Due to the symmetry of *Y*, sampling on the middle 50% yield estimates from NMBE closer to the true values. 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 populations generated from 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 from the middle 50% resulted to 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

Table 6 summarizes the average absolute percent differences of both design-unbiased (SRS) and NMBE estimates from the true population total. The percentage advantage of NMBE (considering the entire population) over SRS is summarized in Table 7.

Table 6. Average Absolute Percent Difference of the Estimates

Variance of X	Sampling Rate														
	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

Variance of X	Sampling Rate				
	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	80.454
400	81.066	81.124	81.349	82.715	84.131

From Table 6, there are slight changes in the NMBE estimates across different levels of variance of *X*. As expected, estimates of SRS are affected by changes in the variance of X. There is a direct proportional relationship between the variance of the auxiliary and the distance between the estimate for the target variable and the true population values. 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 percentage advantage as variability of the auxiliary variable increases. Increase in the variance of *X* while holding all other simulation settings the same increases the proportion of variability in *f(X)* to the variability of *Y*. This is tantamount to increasing the prediction capability of the auxiliary variable thus resulting to an increase in the advantage of NMBE.

For small 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 there is little variation in *X* and when the *Y* values also do not vary much, sampling only from 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 sample selection 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 significant bias 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* between 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 selection 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

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

Table 8, Average Absolute Percent Difference of the Estimates (Varying Population Size)

Population Size	Sampling Rate														
	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 (Varying Population Size)

Population Size	Sampling Rate				
	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

Estimates from NMBE are robust to population size. Hence, for the same sample size irrespective of the population size will result to similar accuracy of the estimates. 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 further that advantages of NMBE over design-unbiased estimates are also robust to population size.

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

5.4 Effect of Model Fit

We presented in Table 10 the average absolute percent differences of both design-unbiased (SRS) and NMBE estimates from the true population total for different levels of model fit. The percentage advantage of NMBE estimates (using the entire population) over SRS estimates are shown in Table 11.

Table 10
Average Absolute Percent Difference of the Estimates (By Value of the Error Multiplier, *k)*

Error Term Multiplier	Sampling Rate														
	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 (Good Fit)	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	2.381	2.611	0.646	1.999	1.975	0.480	1.846	1.367	0.324	1.631	0.851	0.206	1.358
>8.75 (Poor Fit)	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 (By Value of the Error Multiplier, *k)*

Error Term Multiplier	Sampling Rate				
	1%	3%	5%	10%	20%
1 (Good Fit)	87.175	87.043	87.230	88.101	89.259
5	75.687	75.259	75.696	76.298	75.793
>8.75 (Poor Fit)	64.082	63.141	64.013	64.092	62.244

The average absolute percent difference of both SRS and NMBE estimates vary only minimally with respect to the error multiplier: as *k* increases, the percent difference of SRS and NMBE estimates from the true value also increases. However, the changes in NBME estimates are more pronounced. Note that increasing *k* values decreases the amount of variation in *Y* that is explained by *f(X).* The NMBE estimates are model dependent, thus, its performance is affected by changes in the model fit, thus, NMBE is more advantageous for data generated with good model fit (regardless of the form).

The percent difference of the NMBE estimates from the actual population total increases as the error multiplier increases. Deterioration in the capacity of the auxiliary variable, *X*, to predict values of *Y* will affect accuracy of model-based estimates. 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 of interest. 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 but it decreases as the error multiplier is increased. Model-dependent methods rely on the relationship between variable of interest and auxiliary variable to capture the trends in the data. They would fail as estimation tools if the model, which assumes that such a relationship exists, does not fit the data (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.

Sampling only from the middle 50% of the population, NMBE still performed better across all values of the error multiplier for sampling rate of 10%. The completely random nature of SRS translates to better representation of the population when the sample size is large. NMBE does not depend on the size of the sample but 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.

6. Application in FIES Data

The Family Income and Expenditures Survey (FIES) is conducted by the National Statistics Office every three years. The survey collects data on the income distribution, expenditure behaviour, and saving habits of Filipino households. We use the proposed NMBE in 2003 FIES data with the weighted sample data considered as the population. The variables of interest are annual expenditure on food consumed at the following locations: home (Y_{home}), workplace (Y_{work}), school (Y_{school}), restaurant (Y_{res}), and other places (Y_{out}). The "population" total of these variables is predicted using the total annual expenditure (X_{totex}) with measure of variability CV = 125. The coefficients of variation of the target variables are summarized in Table 12.

Table 12. Coefficient of Variation of Expenditures on Food Consumed

Variable	Y_{home}	Y_{work}	Y_{out}	Y_{school}	Y_{res}
CV	81	293	350	360	942

Shown below are the scatterplots of Y's against total annual expenditure for the sample data with sample size n = 8419 (20% of the total of 42,094 households). There is a linear pattern for the relationship between expenditure on food consumed at home and total expenditure (Fig 1) while Y_{work}, Y_{out}, and Y_{school} exhibit a more random scatter for the same auxiliary variable (Fig 2- Fig 4), indicating the increasing variation in the Y's as the value of X increases. For expenditure on food consumed in restaurants, the behaviour of Y_{res} is apparently constant for small values of total expenditure (Fig 5) where the bulk of the data can be located. The extreme observation (encircled) in Figure 5 is an outlier with respect to X_{totex} only but its Y_{res} is in the range of where the majority of the values of Y_{res} lie. There is an extremely high value of Y_{res} (enclosed in the triangle) which explains the large variability of this variable. The outlying observation (enclosed in the triangle) destroys the fit of the constant function defining the relationship between expenditure on food consumed at restaurants and total expenditure (Fig 5).

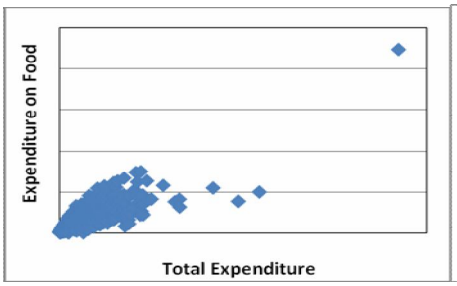

Fig 1. Scatterplot of Total Expenditure vs Expenditure on Food Consumed at Home

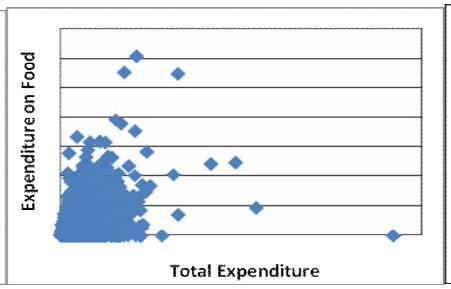

Fig 2. Scatterplot of Total Expenditure vs Expenditure on Food Consumed at the Workplace

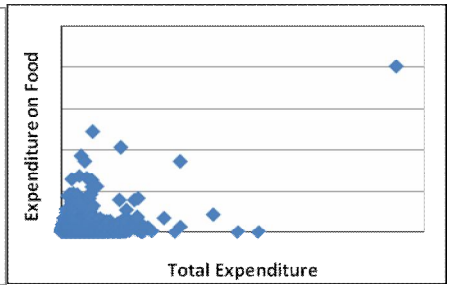

Fig 3. Scatterplot of Total Expenditure vs Expenditure on Food Consumed Outside the Home

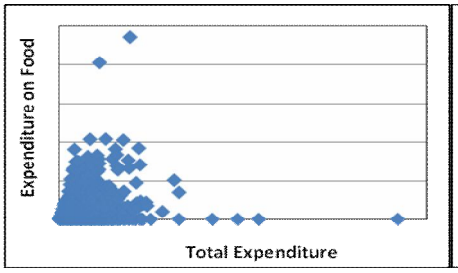

Fig 4. Scatterplot of Total Expenditure vs Expenditure on Food Consumed in School

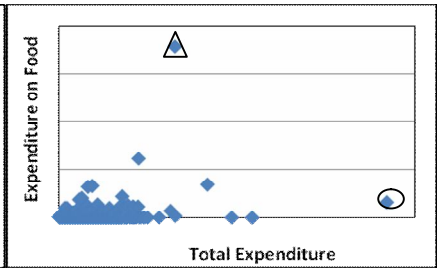

Fig 5. Scatterplot of Total Expenditure vs Expenditure on Food Consumed in Restaurants

Table 13. Average Absolute Percent Difference of the Estimates

	1%		3%		5%		10%		20%	
	SRS	NMBE	SRS	NMBE	SRS	NMBE	SRS	NMBE	SRS	NMBE
Y_{home}	3.228	1.801	1.666	1.122	1.293	0.923	0.921	0.659	0.629	0.438
Y_{work}	10.044	12.825	5.787	5.433	4.656	5.610	2.940	3.424	2.044	9.856
Y_{out}	13.450	19.551	7.394	8.152	6.128	6.456	3.838	4.473	2.319	58.312
Y_{school}	13.926	16.515	8.711	11.049	5.484	7.010	4.426	5.385	3.037	14.034
Y_{res}	28.183	46.632	17.696	19.993	15.193	24.148	11.976	13.122	8.466	20.609

The variables in the Table 13 above are arranged according to increasing coefficient of variation. It can be observed that the average absolute percent difference of SRS estimates are also increasing. This indicates the reduced efficiency of SRS-unbiased estimators as the population is made more heterogeneous. However, they perform better as the sample size is increased.

Among the target variables, of Y_{home}, satisfies the requirement under which, NMBE is superior to design-unbiased estimation. Thus, for all the variables except Y_{home}, SRS-unbiased estimates have lower average percent differences from the true population total than NMBE estimates. It should be noted NMBE for variables with no evident functional pattern can be seen in the scatterplots are either inferior of comparable to design-unbiased estiamtes.

The percent differences of NMBE estimates decrease as the sample size is increased but this trend is observed only up to the 10% sampling rate. The percent difference of NMBE for the 20% sampling rate is higher than that for some of the

other sampling rates. Rao (2005) noted that the efficiency of model-dependent methods can suffer when large samples are used if the model is not correctly specified. Forcing model-based techniques even with a nonparametric nature when no relationship exists is equivalent to model-misspecification. Also, the condition of constant variance for the error term, which was imposed in the simulations, is not satisfied by some variables of the FIES data as shown in the scatterplots.

The variable Y_{home} and total expenditure has a Pearson correlation coefficient of 0.83 and this linear pattern is exhibited in Figure 1. The average absolute percent difference of the NMBE estimate for this case is notably lower given a 1%. The two estimators exhibited the same performance for the other sampling rates.

7. Conclusions

We proposed a nonparametric model-based estimation (NMBE) procedure in estimating population characteristics from survey data. The intensive simulation study illustrates the advantage of NMBE estimates over design-unbiased estimates for a wide variety of simulation settings. NMBE estimates are better than SRS estimates in population with high coefficient of variation and high proportion of variation in *Y* that is accounted for by *X.* Moreover, a pattern in the relationship between the target variable the auxiliary variable must be present for the proposed estimator to be optimal. In the absence of a pattern, the proposed method will only be as good as the design-unbiased method. The method is fairly robust to the form of the data-generating model.

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 (2009) 24:1 – 14

Rao, J. (2005). *Interplay Between Sample Survey Theory and Practice: An Appraisal*. Statistics Canada 31: 117 – 138

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

Xiang, D. (2001). *Fitting Generalized Additive Models with the GAM Procedure.* SAS SUGI 26 Proceeding: P256-26

SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.

Lohr, S. (1999). *Sampling: Design and Analysis.* Brooks/Cole.

Eubank, R. (1998). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, Inc.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models, 1st ed.* Chapman and Hall

Mooney, C. and Duval, R. (1993). *BOOTSTRAPPING A Nonparametric Approach to Statistical Analysis*. Sage Publications, Inc.

StatSoft, Inc. (2010). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/