Two Levels Model Calibration in Cluster Sampling; Use of Penalized Splines in Semiparametric Estimation

Pius Nderitu Kihara[1*], Romanuse Otieno Odhiambo[2], John Kihoro[3]
1. Department of Statistics and Actuarial Sciences, Technical University of Kenya
2. Department of Statistics and Actuarial Sciences, JKUAT, Kenya
3. Department of Distance Learning, Cooperative University College, Kenya
*Email of the correspondiing author: piuskihara@yahoo.com

Abstract

Estimation of finite population total using internal calibration and model assistance on semiparametric models based on kernel methods have been considered by several authors. In this paper, we have extended this to consider model calibration based on penalized splines in two stage sampling where the auxiliary information is available both at the element level and at the cluster level. We have shown that the proposed estimators are robust in the face of misspecified models, are asymptotic design unbiased, have reduced model bias, are consistent and asymptotic normal. We have shown that estimators based on penalized splines perform better than corresponding kernel based estimators and model calibrated estimators perform better than internally calibrated estimators do. .

Keywords: model assistance, model calibration, semiparametric model, penalized splines

1. Introduction

Use of nonparametric and semiparametric modeling techniques for the missing values has gained popularity due to the failings of parametric modeling when a model is misspecified. Given a sample s of n triple of observations $(Z_i, x_i, y_i), i = 1,2,...., n$ from a population U of size N, of interest is to find an estimator for $E(y_i) = g(x_i, Z_i)$ of a missing population value. The auxiliary information consists of a single univariate nonparametric term x and a parametric vector Z composed of an arbitrary number of linear terms. Once the missing values are imputed, an estimate of the population total of the dependent variable Y can be obtained. Breidt et al (2007) [4] considered a super population regression model, ξ given by

$$E_{\xi}(y_i) = g(x_i, Z_i) = \mu(x_i) + Z_i\beta \tag{1}$$

and used a sample estimate of the form $\hat{g}_i = \hat{\mu}(x_i) + Z_i\hat{\beta}$ with $\hat{\mu}(x_i)$ obtained by local polynomial nonparametric method. Accordingly, they obtained the following estimator for population total

$$y_{reg} = \sum_U \hat{g}_i + \sum_s \frac{y_i - \hat{g}_i}{\pi_i} \tag{2}$$

They found that the estimator shares some desirable properties with the fully parametric regression estimators. It is location and scale invariant, and it is internally calibrated for both the parametric and the nonparametric components, in the sense that $\hat{X}_{reg} = \sum_U x_i$ and $\hat{Z}_{reg} = \sum_U Z_i$. The estimator was shown to be design consistent with the rate $\sqrt{n}$, in the sense that $y_{reg} = \sum_U y_i + O_p(\frac{1}{\sqrt{n}})$
In this paper, we extend the work of Breidt et al (2007) [4] to include model calibration in two stage sampling with auxiliary information available at both element and cluster levels.

2. Two Level Model Calibration in Two Stage Survey Sampling

Consider a population U partitioned into M clusters each of size N_i so that the population of clusters is $C = 1,...,i,...,M$. For all clusters $i \in s$, an auxiliary vector x_i and a categorical vector z_i are available.

For simplicity, we let x_i be a scalar. At stage one, a probability sample s of clusters is drawn from C according to a fixed design $p_1(.)$, where $p_1(s)$ is the probability of drawing the sample s from C. let m be the size of s. The cluster inclusion probabilities $\pi_i = p(i \in s)$ and $\pi_{ij} = p(i, j \in s)$ are assumed to be strictly positive and p_1 refers to first stage design. From every sampled cluster $i \in s$, a probability sample s_i of elements is drawn according to a fixed size design $p_1(.)$ with inclusion probabilities $\pi_{k/i} = p(k \in s_i / i \in s)$ and $\pi_{kl/i} = p(k,l \in s_i / i \in s)$. We let n_i be the size of s_i and assume invariance and independence of the second stage design. Let $t_i = g(x_i, Z_i) + \varepsilon_i, \quad i = 1,2,...,M$ where $g(x_i, Z_i)$ is a smooth function of x and Z be the fitted model mean for the ith cluster total. Let $\hat{t}_s = \left[\hat{t}_i\right]_{i\in s}$ be the m vector of $\hat{t}_i$ obtained in the sample of clusters.

Now, consider the case where there is also auxiliary information is known at element level such that for each element in the ith cluster, a nonparametric variable x_{ik} and a categorical vector Z_{ik} are available. Suppose not all elements in a given cluster are available and have to be imputed, we derive a model calibrated estimator of cluster total making use of auxiliary information available at the element level and using penalized splines. Let X_{ci} represents the matrix with rows

$$X_{cik}^T = \left\{1, x_{ik}, ..., x_{ik}^q (x_{ik} - k_1)_+^q, ..., (x_{ik} - k_\kappa)_+^q\right\} \tag{3}$$

for $i \in C_i$, and let Y denote the column vector of response values y_{ik} for $k \in C_i$ so that $\hat{y}_{si} = \left[\hat{y}_{ik}\right]_{k\in s_i}$ be the vector of $\hat{y}_{ik}$ obtained in the sample of cluster .

Let $A_\alpha = diag\{0,...,0,\alpha\}$, with q+1 zeros on the diagonal followed by k penalty constants α. We adapt the definition of the matrix of inverse inclusion probabilities by Breidt et al (2005) [1] to the matrix of within cluster inclusion probabilities as $w_{si} = diag_{k\in s_i}\left(\pi_{k/i}^{-1}\right)$. Let X_{cisi} be the sub matrix of X_{ci} consisting of those rows for which $k \in s_i$

Let ξ_{11} denote the superpopulaton of cluster elements model. We define the semiparametric population estimator for $\mathrm{E}_{\xi_{11}}\left(y_{ik}\right)$ as

$$\hat{g}_{ik} = \hat{g}\ (x_{ik}, z_{ik}) = \hat{\mu}\ (x_{ik}) + Z_{ik}\hat{\beta}_i \tag{4}$$

and design weighted penalized spline smoother vector be

$$Ss_{ik} = \left(X^T{}_{cisi} W_{si} X_{cisi} + A_\alpha\right) X_{cisi} W_{si} \tag{5}$$

The sample smoother matrix is given by the following.

$$S_{si} = \left[S_{sik}, k \in s_i\right] \tag{6}$$

Accordingly, we have the following estimators resulting from the solution of the equations (4), (5), and (6).

$$\hat{\beta}_i = \left(Z^T{}_{si} S_{si} Z_{si} + A_\alpha\right)^{-1} Z^T{}_{si} S_{si} \hat{y}_{si} \tag{7}$$

$$\hat{\mu}_{ik} = \hat{\mu}(x_{ik}) = S_{sik}(\hat{y}_{si} - Z^T{}_{si}\hat{\beta}_i) \tag{8}$$

Where $\hat{\mu}_{ik}$ and x_{ik} are defined for every $k \in C_i$. We propose a semiparametric model assisted model calibrated estimator of cluster total to be

$$\hat{t}_i = \sum_{k\in s_i} w_{ik}\hat{y}_{ik} \tag{9}$$

With w_{ik} obtained by minimizing the chi square distance measure

$$\Phi_s = \sum_{k\in s_i} \frac{\left(w_{ik} - d_{ik}\right)^2}{q_{ik} d_{ik}} \tag{10}$$

Subject to the constraints $\sum_{k\in s_i} w_{ik} = N_i$ and $\sum_{k\in s_i} w_{ik}\hat{g}_{ik} = \sum_{k\in C_i} \hat{g}_{ik} = N_i$ which we adopt from constraints

introduced by Wu and Sitter (2001) [8]. Here, $d_{ik} = \pi^{-1}{}_{k/i}$ and q_{ik} are known positive constants uncorrelated with the d_{ik} . See Deville and Sarndal, (1992) [5].

We introduce the langrage procedure in the minimization of equation (10) obtain the equation below.

$$l = \sum_{k\in s_i} \frac{(w_{ik-d_{ik}})^2}{q_{ik} d_{ik}} - 2\lambda\Big(\sum_{k\in s_i} w_{ik}\hat{g}_{ik} - \sum_{k\in C_i} \hat{g}_{ik}\Big) - 2v\Big(\sum_{k\in s_i} w_{ik} - N_i\Big) \tag{11}$$

where λ is the langrage's multiplier and v is the penalty constant. Differentiating l with respect to w_{ik} , equating the derivative zero and solving we get

$$w_{ik} = (\lambda \hat{g}_{ik} + v) q_{ik} d_{ik} + d_{ik} \tag{12}$$

Solving for λ and v , and substituting in $\hat{t}_i$ we have that

$$\hat{t}_i = \sum_{k\in s_i} d_{ik}\hat{y}_{ik} + \Big(M - \sum_{k\in s_i} d_{ik}\Big)\left\{\frac{\sum_{k\in s_i} d_{ik} q_{ik}\hat{y}_{ik}}{\sum_{k\in s_i} d_{ik} q_{ik}} - \hat{\beta}_{mc}\right\} + \left\{\sum_{k\in C_i}\hat{g}_{ik} - \sum_{k\in s_i} d_{ik}\hat{g}_{ik}\right\}\hat{\beta}_{mc} \tag{13}$$

where $$\hat{\beta}_{mc} = \left\{\frac{\sum_{k\in s_i} q_{ik} d_{ik}\left(\hat{g}_{ik} - \frac{\sum_{k\in s_i} d_{ik} q_{ik}\hat{g}_{ik}}{\sum_{k\in s_i} d_{ik} q_{ik}}\right)\left(\hat{y}_{ik} - \frac{\sum_{k\in s_i} d_{ik} q_{ik}\hat{y}_{ik}}{\sum_{k\in s_i} d_{ik} q_{ik}}\right)}{\sum_{k\in s_i} q_{ik} d_{ik}\left(\hat{g}_{ik} - \frac{\sum_{k\in s_i} d_{ik} q_{ik}\hat{g}_{ik}}{\sum_{k\in s_i} d_{ik} q_{ik}}\right)^2}\right\}$$

The term $\Big(M - \sum_{k\in s_i} d_{ik}\Big)\left\{\frac{\sum_{k\in s_i} d_{ik} q_{ik}\hat{y}_{ik}}{\sum_{k\in s_i} d_{ik} q_{ik}} - \hat{\beta}_{mc}\right\}$ has been shown from empirical analysis to be negligible

and has no effect on asymptotic properties hence we rewrite the estimator as

$$\hat{t}_i = \sum_{k\in s_i} \frac{\hat{y}_{ik}}{\pi_{k/i}} + \left\{\sum_{k\in C_i}\hat{g}_{ik} - \sum_{k\in s_i}\frac{\hat{g}_{ik}}{\pi_{k/i}}\right\}\hat{\beta}_{mc} \tag{14}$$

Now, having estimated the cluster totals, we then derive an estimator of the population total using the estimated cluster totals and the auxiliary information available at cluster level. Define the spline model matrix X_c to contain bases that are functions of $\hat{t}_i$ and define the sub matrix $W_s = diag_{j\in s}\left(\pi_j^{-1}\right)$. Let ξ_1 denote the super population of clusters model. Define the semiparametric population estimator for $\mathrm{E}_{\xi_1}\left(\hat{t}_i\right)$ as

$$\hat{g}_i = \hat{g}\ (x_i, z_i) = \hat{\mu}\ (x_i) + Z_i\hat{\beta} \tag{15}$$

and design weighted penalized spline smoother vector be

$$S_{si} = \left(X^T{}_{cs} W_s X\ {}_{cs} + A_\alpha \right) X_{cs} W\ {}_s \tag{16}$$

while the sample smoother matrix is given by

$$S_s = \left[S_{si}, i \in s \right] \tag{17}$$

Again, we have the following estimators resulting from the solution of the equations (15), (16) and (17).

$$\hat{\beta} = \left(Z^T{}_s S_s Z_s + A_\alpha \right)^{-1} Z^T{}_s S_s \hat{t}_s \tag{18}$$

$$\hat{\mu}_i = \hat{\mu}(x_i) = S_{si}(\hat{t}_s - Z^T{}_s \hat{\beta}) \tag{19}$$

With $\hat{\mu}_i$ and x_i defined for every $i \in U$. We propose a semparametric model assisted model calibrated estimator of population total as

$$\hat{y}_{sm2} = \sum_{i \in s} w_i \hat{t}_i \tag{20}$$

with w_i obtained by minimizing the chi square distance measure

$$\Phi_s = \sum_{i \in s} \frac{\left(w_i - d_i \right)^2}{q_i d_i} \tag{21}$$

Subject to the constraints $\sum_{k \in s} w_i = N$ and $\sum_{i \in s} w_i \hat{g}_i = \sum_{i \in U} \hat{g}_i$. Again, $d_i = \pi^{-1}{}_i$ and q_i are known

positive constants uncorrelated with d_i.We introduce the langrage procedure in the minimization of equation (21)

to obtain the following estimator of population total

$$\hat{y}_{sm2} = \sum_{i \in s} d_i \hat{t}_i + (M - \sum_{i \in s} d_i) \left\{ \frac{\sum_{i \in s} d_i q_i \hat{t}_i}{\sum_{i \in s} d_i q_i} - \hat{\beta}_m \right\} + \left\{ \sum_{i \in U} \hat{g}_i - \sum_{i \in s} d_i \hat{g}_i \right\} \hat{\beta}_m \tag{22}$$

Where $\hat{\beta}_m = \left\{ \dfrac{\sum_{i \in s} q_i d_i \left(\hat{g}_i - \dfrac{\sum_{i \in s} d_i q_i \hat{g}_i}{\sum_{i \in s} d_i q_i} \right) \left(\hat{t}_i - \dfrac{\sum_{i \in s} d_i q_i \hat{t}_i}{\sum_{i \in s} d_i q_i} \right)}{\sum_{i \in s} q_i d_i \left(\hat{g}_i - \dfrac{\sum_{i \in s} d_i q_i \hat{g}_i}{\sum_{i \in s} d_i q_i} \right)^2} \right\}$

The term $(M - \sum_{i \in s} d_i) \left\{ \dfrac{\sum_{i \in s} d_i q_i \hat{t}_i}{\sum_{i \in s} d_i q_i} - \hat{\beta}_m \right\}$ is again negligible so we rewrite as

$$\hat{y}_{sm2} = \sum_{i\in s}\frac{\hat{t}_i}{\pi_i} + \left\{\sum_{i\in U}\hat{g}_i - \sum_{i\in s}\frac{\hat{g}_i}{\pi_i}\right\}\hat{\beta}_m \tag{23}$$

A corresponding internally calibrated estimator will therefore be

$$\hat{y}_{reg2} = \sum_{i\in s}\frac{\hat{t}_i}{\pi_i} + \left\{\sum_{i\in U}\hat{g}_i - \sum_{i\in s}\frac{\hat{g}_i}{\pi_i}\right\} \tag{24}$$

We now derive the variance of the population estimator (23). If the sample comprises the whole population of clusters, then $\hat{y}_{sm2} = \sum_{i\in s}\frac{t_i}{\pi_i}$ which is the Horvtz -Thompson (HT) design based estimator and as shown by Breidt et al (2005) [2],

$$\mathrm{var}_p(\hat{y}_{sm2}) = V_1(\mathrm{E}_{11}[\hat{y}_{sm2}]) + \mathrm{E}_1(V_{11}[\hat{y}_{sm2}]) \tag{25}$$

$$= \sum_{i\in C}\sum_{j\in C}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{t_i}{\pi_i}\frac{t_j}{\pi_j} + \sum_{i\in C}\frac{V_i}{\pi_i} \tag{26}$$

Now, the variance component at the element level within a cluster is

$V_i = V_{11}(\hat{t}_i) = \sum_{k}^{m}\sum_{l}^{m}\left(\pi_{kl/i} - \pi_{k/i}\pi_{l/i}\right)\frac{y_{ik} - \hat{g}_{ik}\hat{\beta}_{mc}}{\pi_{k/i}}\frac{y_{il} - \hat{g}_{il}\hat{\beta}_{mc}}{\pi_{l/i}}$ due to the presence of the model

component $\left\{\sum_{k=1}^{Ni}\hat{g}_{ik} - \sum_{k\in s_i}d_{ik}\hat{g}_{ik}\right\}\hat{\beta}_{mc}$. When $\hat{y}_{sm2}$ has the model component $\left\{\sum_{i=1}^{M}\hat{g}_i - \sum_{i=1}^{m}d_i\hat{g}_i\right\}\hat{\beta}_m$,

its design variance becomes

$$\sum_{i\in C}\sum_{j\in C}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{t_i - \hat{g}_i\hat{\beta}_m}{\pi_i}\frac{t_j - \hat{g}_j\hat{\beta}_m}{\pi_j} + \sum_{i\in C}\frac{V_i}{\pi_i} \tag{27}$$

3. Asymptotic Properties

We now establish the asymptotic properties for $\hat{y}_{sm2}$

3.1 Assumptions

1. We assume that there is a sequence of finite populations indexed by ρ each of size N_ρ but which we compress and write N .
2. As $\rho \to \infty, N, n, M, m, N_i, n_i \to \infty$. Also, the number of knots $k \to \infty$ while bandwidth $h \to 0$.
3. For each ρ , the x_i , $i = 1, 2, ..., M$ are independent and identically distributed $F(x) = \int_{-\infty}^{x} g(t)dt$ where $g(.)$ is a density with compact support $[a_x, b_x]$ and $g(x) > 0$ for all $x \in [a_x, b_x]$. The Z_i has bounded support.
4. For each ρ , the x_i are considered fixed with respect to the model ξ_1 while the errors ε_{i1} are independent and have mean zero, variance $\mathrm{var}(x_i, Z_i)$ and compact support, uniformly for each ρ .
5. For each ρ , the x_{ik} are considered fixed with respect to the model ξ_{11} while the errors ε_{i11} are independent and have mean zero, variance $\mathrm{var}(x_{ik}, Z_{ik})$ and compact support, uniformly for each ρ .
6. The sampling design is regular so that the inclusion probabilities are independent of response measurements and satisfies the following conditions ;

   a) $\max_{i\in s}\frac{m}{M\pi_i} = 0(1), and \max_{k\in s_i}\frac{n_i}{N_i\pi_{k/i}} = 0(1)$

b) $\sum_{i\in s}\frac{g_i}{\pi_i}-\sum_{i=1}^{M} g_i = o_p(Mm^{-\frac{1}{2}}), and \sum_{k\in s_i}\frac{g_{ik}}{\pi_{k/i}}-\sum_{k=1}^{N_i} g_{ik} = o_p(N_i n_i^{-\frac{1}{2}})$

First condition says that no basic design weight is disproportionally large while the second condition is equivalent to assuming that Horvitz Thompson estimators for $\sum_{i=1}^{M} g_i$ and $\sum_{k=1}^{N_i} g_{ik}$ are asymptotically normally distributed.

7. Let g_i be the population fit and $\hat{\hat{y}}_{sm2}=\sum_{i=1}^{m}\frac{\hat{t}_i}{\pi_i}+\left\{\sum_{i=1}^{M} g_i-\sum_{i=1}^{m}\frac{g_i}{\pi_i}\right\}\hat{\hat{\beta}}_m$ where

$\hat{\hat{\beta}}_m=\frac{\sum_{j=1}^{M}\frac{1}{\pi_i}q_i(g_i-\bar{g})(\hat{t}_i-\bar{t})}{\sum_{i=1}^{M}\frac{1}{\pi_i}q_i(g_i-\bar{g})^2}$ and $\bar{g}=\sum_{i=1}^{M} g_i$

Under a regular sampling design (assumption 6), $Avar(\hat{y}_{sm2})=var(\hat{\hat{y}}_{sm2})$. The variance of the asymptotic distribution of $\hat{y}_{sm2}$ can therefore be consistently estimated mild assumptions.

3.2 Asymptotic Design Unbiasedness

Let E_{p_1} be design expectation and E_{ξ_1} model based expectation. We need to show that $E_{p_1}(\hat{y}_{sm2})=Y_t$. We note that $\hat{t}_i$ is a Horvitz Thompson design estimator which is unbiased fot t_i . Now,

$$\mathrm{E}_{p_1}\left(\hat{y}_{sm2}\right)=\mathrm{E}_{p_1}\left\{\sum_{i\in s}\frac{\hat{t}_i}{\pi_i}+\left\{\sum_{i\in U}\hat{g}_i-\sum_{i\in s}\frac{\hat{g}_i}{\pi_i}\right\}\hat{\beta}_{m2}\right\} \tag{28}$$

$$=\mathrm{E}_{p_1}\left\{\sum_{i\in U}\frac{\hat{t}_i I_i}{\pi_i}+\left\{\sum_{i\in U}\hat{g}_i-\sum_{i\in U}\frac{\hat{g}_i I_i}{\pi_i}\right\}\hat{\beta}_{m2}\right\} \tag{29}$$

$$=\left\{\sum_{i\in U}\frac{\mathrm{E}_{p_1}\hat{t}_i I_i}{\pi_i}+\left\{\sum_{i\in U}\mathrm{E}_{p_1}\hat{g}_i-\sum_{i\in U}\frac{\mathrm{E}_{p_1}\hat{g}_i I_i}{\pi_i}\right\}\mathrm{E}_{p_1}\hat{\beta}_{m2}\right\} \tag{30}$$

$$==\left\{\sum_{i\in U}\frac{t_i}{1}+\left\{\sum_{i\in U}\hat{g}_i-\sum_{i\in U}\frac{\hat{g}_i}{1}\right\}\mathrm{E}_{p_1}\hat{\beta}_{m2}\right\} \tag{31}$$

Since $\mathrm{E}p_1\left(I_i\right)=\pi_i$ and with respect to design expectation, $\hat{g}_i$ is treated as a constant. Thus, we have $\sum_{i\in U} t_i = Y_t$.

3.3 Model Bias Reduction

$\hat{\beta}_m$ is an estimate of the change in Y_t when g_i is increased by a unit. If $\sum_{i\in s}\frac{\hat{g}_i}{\pi_i}$ is below average,

we should expect the population total Y_t to be below average by an amount $\left\{\sum_{i\in U}\hat{g}_i-\sum_{i\in s}\frac{\hat{g}_i}{\pi_i}\right\}\hat{\beta}_m$ due to regression of $\hat{t}_i$ on $\hat{g}_i$. See Cochran (1997) [4]. Again, the estimate $\hat{g}_i$ need not be free from bias. If $\hat{g}_i-\hat{t}_i=D$, so that the estimate is perfect except for a constant bias D, then with $\hat{\beta}_m=1$ the regression estimate becomes

$$\sum_{i\in s}\frac{\hat{t}_i}{\pi_i}+\left\{\sum_{i\in U}\hat{g}_i-\sum_{i\in s}\frac{\hat{g}_i}{\pi_i}\right\}=\sum_{i\in U}\hat{g}_i+\left\{\sum_{i\in s}\frac{\hat{t}_i}{\pi_i}-\sum_{i\in s}\frac{\hat{g}_i}{\pi_i}\right\} \tag{32}$$

=Population total estimates + adjustment for bias.

This regression estimate is consistent in the sense that when the sample comprises the whole population, then $\sum_{i\in U}\hat{g}_i=\sum_{i\in s}\frac{\hat{g}_i}{\pi_i}$ and the regression estimate reduces to $\sum_{i\in s}\frac{\hat{t}_i}{\pi_i}$. See Firth and Bennett (2006) [6]. Again, establishing a CLT for $\hat{y}_{sm2}$, which is a generalized difference estimator is essentially the same as establishing a CLT for Horvitz-Thompson estimator.

3.4 Design Consistency

Using chebycheve's inequality and a sequence of the estimates $\hat{y}_{sm2\rho}$ but which we compress to $\hat{y}_{sm2}$,
We have that $pr[|\hat{y}_{sm2}-Y_t|>\varepsilon]\le E_{P_1}\frac{|\hat{y}_{sm2}-Y_t|^2}{\varepsilon^2}$
But since $\hat{y}_{sm2}$ is unbiased for Y_t, then the mean squared error is consistently estimated by $var(\hat{y}_{sm2})$,
so that $pr[|\hat{y}_{sm2}-Y_t|>\varepsilon]\le\frac{var\{\hat{y}_{sm2}\}}{\varepsilon^2}$

and $\lim_{\rho\to\infty}pr[|E_{p_1}\hat{y}_{sm2}-Y_t|>\varepsilon]\le\lim_{\rho\to\infty}\frac{var\{\hat{y}_{sm2}\}}{\varepsilon^2}$. We see that,

$$\begin{aligned}\lim_{\rho\to\infty}\frac{var\{\hat{y}_{sm2}\}}{\varepsilon^2}&=\lim_{\rho\to\infty}\mathrm{E}_{p1}\sum_{i=1}^{m}\sum_{i=1}^{m}\frac{t_i-\hat{g}_i\hat{\beta}_{m2}}{\pi_i}\frac{t_j-\hat{g}_j\hat{\beta}_{m2}}{\pi_j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{1}{\varepsilon^2}+\\&\lim_{\rho\to\infty}\mathrm{E}_{p1}\sum_{i=1}^{M}\left\{\sum_{k=1}^{m}\sum_{l=1}^{m}(\pi_{kl/i}-\pi_{k/i}\pi_{l/i})\frac{y_k}{\pi_{k/i}}\frac{y_l}{\pi_{l/i}}\right\}\frac{1}{\varepsilon^2\pi_i}\end{aligned} \tag{33}$$

$$\begin{aligned}&=\lim_{\rho\to\infty}\mathrm{E}_{p1}\sum_{i=1}^{M}\sum_{i=1}^{M}\frac{t_i-\hat{g}_i\hat{\beta}_{m2}}{\pi_i}\frac{t_j-\hat{g}_j\hat{\beta}_{m2}}{\pi_j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{\mathrm{I}_i\mathrm{I}_j}{\varepsilon^2}+\\&\lim_{\rho\to\infty}\mathrm{E}_{p1}\sum_{k=1}^{M}\sum_{l=1}^{M}(\pi_{kl/i}-\pi_{k/i}\pi_{l/i})\frac{y_k}{\pi_{k/i}}\frac{y_l}{\pi_{l/i}}\frac{\mathrm{I}_i\mathrm{I}_j}{\varepsilon^2\pi_i}=0\end{aligned} \tag{34}$$

since $E_{p_1}(\pi_{ij})=\pi_i\pi_j, E_{p_1}(\pi_i\pi_j)=\pi_i\pi_j, E_{p_1}(\pi_i)=\pi_i$ and $E_{p_1}(I_iI_j)=\pi_{ij}\le\pi_i\pi_j$.
Therefore, $\lim_{\rho\to\infty}pr[|E_{p_1}\hat{y}_{sm2}-Y_t|>\varepsilon]\to 0$. That is , $\hat{y}_{sm2}\xrightarrow{p}Y_t$

3.5 Asymptotic Normality

<u>Theorem 1</u>: Let $\hat{\hat{y}}_{sm2}$ be as defined in assumption 7. Then,

$\frac{M^{-1}(\hat{\hat{y}}_{sm2}-Y_t)}{var^{1/2}(M^{-1}\hat{\hat{y}}_{sm2})}\to N(0,1)$ as $\rho\to\infty$ implies that $\frac{M^{-1}(\hat{y}_{sm2}-Y_t)}{var^{1/2}(M^{-1}\hat{y}_{sm2})}\to N(0,1)$

where

$$var(M^{-1}y\hat{}_{sm2})=\frac{1}{M^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{\hat{t}_i-\hat{g}_i\hat{\beta}_{m2}}{\pi_i}\right)\left(\frac{\hat{t}_j-\hat{g}_j\hat{\beta}_{m2}}{\pi_j}\right)\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right) \tag{35}$$

<u>Proof</u>: We need to show that $(\hat{y}_{sm2}-Y_t)$ converges to $(\hat{\hat{y}}_{sm2}-Y_t)$ in distribution. This would imply that

$\hat{y}_{sm2}$ inherits limiting distributional properties of $\hat{\hat{y}}_{sm2}$. This, coupled by assumption 7 would proof the above. Now,

$$(\hat{y}_{sm2} - Y_t) = \sum_{i=1}^{M} \frac{t\hat{}_i I_i}{\pi_i} + \sum_{i=1}^{M} g\hat{}_i \hat{\beta}_{m2} - \sum_{i=1}^{M} \frac{\hat{g}_i I_i \hat{\beta}_{m2}}{\pi_i} - \sum_{i=1}^{M} \hat{t}_i$$

and $(\hat{\hat{y}}_{sm2} - Y_t) = \sum_{i=1}^{M} \frac{t\hat{}_i I_i}{\pi_i} + \sum_{i=1}^{M} g_i \hat{\hat{\beta}}_{m2} - \sum_{i=1}^{M} \frac{g_i I_i \hat{\hat{\beta}}_{m2}}{\pi_i} - \sum_{i=1}^{M} \hat{t}_i$. Clearly,

$$\hat{y}_{sm2} - \hat{\hat{y}}_{sm2} = \sum_{i=1}^{M} \left(\hat{g}_i \hat{\beta}_{m2} - g_i \hat{\hat{\beta}}_{m2} \right) \left(1 - \frac{I_i}{\pi_i} \right) \tag{36}$$

Taking limits of the expectation, we have

$$\lim_{\rho \to \infty} E_{P_1} \left\{ y\hat{}_{sm2} - \hat{\hat{y}}_{sm2} \right\} = \lim_{\rho \to \infty} E_{P_1} \left\{ \sum_{i=1}^{M} \left(\hat{g}_i \hat{\beta}_{m2} - g_i \hat{\hat{\beta}}_{m2} \right) \left(1 - \frac{I_i}{\pi_i} \right) ht \right\} \tag{37}$$

It can be seen that the design expectation of $\hat{y}_{sm2} - \hat{\hat{y}}_{sm2}$ approaches zero since design expectation of I_i is π_i. This is convergence in mean which implies convergence in probability and convergence in distribution.

4. Empirical Analysis

We simulated a population of independent and identically distributed variable x using uniform (0.1) and a categorical matrix Z. For each generated x_i and vector Z_i and for each mean function, $N_i = 100$ element values were generated as follows.

$$y_{ik} = \frac{g(x_i, Z_i)}{\sqrt{N_i}} + \frac{\varepsilon_{ik}}{\sqrt{N_i}}, \{\varepsilon_{ik}\} iid N(0, 0.1) \tag{38}$$

where y_{ik} is the kth element in the ith cluster and $g(x_i, Z_i)$, which we simply write g_i is the mean function for the cluster total t_i. This generating function is an adaptation to semiparametric modeling of the generating function by Montanari and Ranalli (2006) [7].

We considered the following mean functions for auxiliary information at cluster level.

1. $linear \quad Z\beta' + 2 + 5x$
2. $quadratic \quad Z\beta' + (2+5x)^2$
3. $bump \quad Z\beta' + (2+5x) + \exp(-200(2+5x)^2)$
4. $\exp onential \quad Z\beta' + \exp(-8x)$
5. $cycle \quad 1 \quad Z\beta' + \sin e(2\pi x)$
6. $cycle \quad 2 \quad Z\beta' + \sin e(8\pi x)$

For simplicity, within each cluster, the auxiliary information x_{ik} at element level was generated using the linear and quadratic mean functions and working backward to obtain the following respective formulas.

$$x_{ik} = \frac{y_{ik} - 2 - z_{ik}\beta'}{5} \tag{39}$$

and

$$x_{ik} = \frac{-2 + \sqrt{y_{ik} - z_{ik}\beta'}}{5} \tag{40}$$

where Z_{ik} is the matrix (Z_{i1}, Z_{i2}, Z_{i3}), Z_{i1} is a matrix of 1s, Z_{i2} is a matrix of 2s, 3s and 4s, while Z_{i3} is a matrix of 5s,6s, and 7s. β is the matrix $(1, 2, 3)$.

For each pair (x_i, Z_i) and mean function, R=100 replicate samples of clusters were generated. At stage one, a sample of clusters was generated by simple random sampling with sample size m=50. At stage two, within each of the selected clusters, sub samples of size $n_i = 50$ were generated by simple random sampling. Where we used penalized splines in fitting a missing cluster element, we also used penalized splines in fitting missing cluster totals, and similarly for local polynomial and Nadaraya Watson kernel methods. Using the estimated cluster totals, estimates of the population total were generated. We compared the performance of several estimators;

1. Horvitz Thompson estimator, $\hat{y}_{ht2}$
2. The model calibrated model assisted semiparametric estimator $\hat{y}_{sm2},(31)$ that we have proposed, for which we considered three cases based on the nonparametric method used to obtain the mean estimate. These are; $\hat{y}_{smsp2}, \hat{y}_{smlp2}, \quad and \quad \hat{y}_{smnw2}$ for penalized splines, local polynomial and Nadaraya Watson kernel smoothing respectively.
3. Internally calibrated model assisted semiparametric estimator $\hat{y}_{reg2},(32)$ for which we consider the three cases; $\hat{y}_{resp2}, \hat{y}_{reglp2}, \quad and \quad \hat{y}_{regnw2}$ for penalized splines, local polynomial and Nadaraya Watson kernel smoothing respectively.

The performance of any estimator say y_{est} in y_{ht2}, $\hat{y}_{smsp2}$, $\hat{y}_{smlp2}$, $\hat{y}_{smnw2}$, $\hat{y}_{regsp2}$, $\hat{y}_{reglp2}$, $\hat{y}_{regnw2}$ was evaluated using its relative bias R_B and relative efficiency R_E defined by

$$R_B = \frac{\sum_{r=1}^{R}\left(y_{est} - Y_t\right)}{R * Y_t} \tag{41}$$

where R is the replicate number of samples and

$$R_E = \frac{MSE(y_{est})}{MSE(\hat{y}_{ht2})} \tag{42}$$

where y_{est} was calculated from the R^{th} simulated sample.
The $\hat{y}_{ht2}$ estimator was used as the baseline comparison. Large values of relative efficiencies,($R_E \geq 1$) represent higher efficiency for $\hat{y}_{ht}$ over y_{est} . We also carried out a Sensitivity Analysis by looking at the effects that ignoring a variable in the categorical matrix would have on the estimators. We dropped values available at cluster level. Same effects would be expected if an auxiliary variable at element level is dropped since the processes of estimation at both stages are similar. We report on the observations for the case where the auxiliary information at element level was generated from the linear function. Similar observations were made when the auxiliary information at the element level was obtained from the quadratic function. Clearly, the results would similarly not be different if any of the six generating functions is considered.

4.1 Bias

Table 1. Absolute Biases

	$\hat{y}_{smsp2}$	$\hat{y}_{smlp2}$	$\hat{y}_{smnw2}$	$\hat{y}_{ht2}$	$\hat{y}_{regsp2}$	$\hat{y}_{reglp2}$	$\hat{y}_{regnw2}$
Linear	0.015	0.015	0.025	0.017	0.028	0.048	0.328
Quadratic	0.041	0.039	0.041	0.039	0.516	1.645	2.906
Bump	0.031	0.036	0.040	0.036	0.048	0.247	0.339
Exponential	0.013	0.016	0.021	0.023	0.014	0.030	0.125
Cycle 1	0.012	0.015	0.023	0.019	0.018	0.034	0.086
Cycle 2	0.012	0.010	0.015	0.022	0.017	0.013	0.028

From table (1), we observe that the biases are very small again pointing to unbiasedness for all the estimators. Comparing each model calibrated estimator with its corresponding internally calibrated estimator, that is, $\hat{y}_{smsp2}$

with $\hat{y}_{regsp2}$, $\hat{y}_{smlp2}$ with $\hat{y}_{reglp2}$ and $\hat{y}_{smnw2}$ with $\hat{y}_{regnw2}$, we see that model calibration results in reduced bias than internal calibration.

4.2 Relative Mean Squared Error

Table 2. Relative Mean Squared Errors

	$\hat{y}_{smsp2}$	$\hat{y}_{smlp2}$	$\hat{y}_{smnw2}$	$\hat{y}_{ht2}$	$\hat{y}_{regsp2}$	$\hat{y}_{reglp2}$	$\hat{y}_{regnw2}$
Linear	1.497	1.242	2.175	1	4.229	8.573	9.004
Quadratic	2.027	2.431	2.730	1	3.933	7.003	10.706
Bump	2.168	2.320	2.743	1	3.454	6.659	8.332
Exponential	2.213	2.630	2.691	1	2.890	5.657	8.553
Cycle 1	2.059	2.641	2.841	1	3.731	6.945	11.077
Cycle 2	2.131	2.172	2.879	1	4.259	7.456	11.321

From table (2), the model calibrated estimators $\hat{y}_{smsp2}$, $\hat{y}_{smlp2}$ and $\hat{y}_{smnw2}$ perform consistently better than the internally calibrated estimators $\hat{y}_{regsp2}$, $\hat{y}_{reglp2}$ and $\hat{y}_{regnw2}$. The penalized spline based model calibrated estimator $\hat{y}_{smsp2}$ performs better than the kernel based model calibrated estimators $\hat{y}_{smlp2}$ and $\hat{y}_{smnw2}$.

4.3 Bias on Sensitivity Analysis

Table 3. Bias on Removing Z_3

	$\hat{y}_{smsp2}$	$\hat{y}_{smlp2}$	$\hat{y}_{smnw2}$	$\hat{y}_{ht2}$	$\hat{y}_{regsp2}$	$\hat{y}_{reglp2}$	$\hat{y}_{regnw2}$
Linear	0.024	0.040	0.040	0.024	0.029	0.302	0.173
Quadratic	0.063	0.092	0.066	0.063	0.067	1.250	0.274
Bump	0.026	0.054	0.041	0.035	0.040	0.431	0.161
Exponential	0.252	0.252	0.253	0.246	0.261	0.710	0.302
Cycle 1	0.024	0.024	0.029	0.022	0.026	0.242	0.063
Cycle 2	0.021	0.022	0.031	0.022	0.028	0.155	0.152

Looking at table (3), we observe that the biases still remain very small even after the variable Z_3 is dropped meaning the estimators still perform well.

4.4 Relative Mean Squared Error on Sensitivity

Table 4. Relative Mean Squared Error on Removing Z_3

	$\hat{y}_{smsp2}$	$\hat{y}_{smlp2}$	$\hat{y}_{smnw2}$	$\hat{y}_{ht2}$	$\hat{y}_{regsp2}$	$\hat{y}_{reglp2}$	$\hat{y}_{regnw2}$
Linear	1.952	2.214	2.897	1	5.112	15.348	19.783
Quadratic	2.017	4.911	5.525	1	5.892	14.006	16.786
Bump	2.022	2.889	3.312	1	4.021	14.134	18.532
Exponential	2.112	2.634	2.992	1	4.289	13.129	19.245
Cycle 1	1.992	2.745	3.429	1	3.987	15.164	18.923
Cycle 2	2.194	3.004	4.101	1	4.934	17.356	19.912

Comparing results of table (2) and table (4), we observe that there is no much change in the efficiency of the model calibrated estimators when Z_3 is dropped. This illustrates the robustness of the model calibrated estimators. For the internally calibrated estimators, there is a noticeable loss of efficiency when Z_3 is dropped.

5. Conclusion

It has been observed that the model calibrated estimators perform better than their corresponding internally calibrated estimators. When penalized splines are used to fit the missing values, the estimators performs well than when local polynomial or Nadaraya Watson smoothing are used. The biases are quite small for all the estimators. It is clear that even the internally calibrated estimators are still reliable.

When some of the categorical variables are not considered in estimation, the model calibrated estimators are found to be more robust than the internally calibrated estimators. In a real world problem where we may not have, or may not be sure that we have all the relevant auxiliary information about a variable, model calibrated estimators would therefore be the estimators of choice.

It is observed that even though using penalized splines results in a more efficient model calibrated or internally calibrated estimator than when kernel based methods are used, an internally calibrated estimator that uses penalized splines is less efficient than a model calibrated estimator that uses kernel based method to fit missing values. Thus, to model calibrate or not is more significant question than the choice of the nonparametric method to use to fit the missing values.

We have shown that in cases where some elements within clusters are unreachable but auxiliary information is available at element level, we can take advantage of this auxiliary information to obtain cluster totals, which are then used in the estimation of population total. We note if there is a possibility that some clusters may be unreachable, it means there is also the possibility that some cluster elements may be unreachable.

References

[1] Breidt, F.J., Claeskens, G. & Opsomer, J.D. (2005), "Model Assisted Estimation for Complex Surveys Using Penalized Splines", *Bometrika*, 92, 831-846.

[2] Breidt, F. J., Kim, J.Y. & Opsomer, J.D. (2005), " Nonparametric Regression Estimation of Finite Population Totals Under Two Stage Sampling", *Annals of Statistics,* 25, 1026-1053.

[3] Breidt, F.J., Opsomer, J.D., Alicia, A.J. & Ranalli, G. (2007), "Semiparametric Model Assisted Estimation for Natural Resource Surveys", *Statistics Canada,* Catalogue No. 12-001.

[4] Cochran W.G. (1997). "Sampling Techniques (3rd ed.)", New York: *John Wiley and sons.*

[5] Deville, J.C. & Sarndal C.E. (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association ,* 87,376-82.

[6] Firth, D. & Bennet, K.E. (2006), "Robust Models in Probability Sampling", *Journal of Royal Statistical Society.* B, 17, 267-278.

[7] Montanari, G.E. & Ranalli, M.G. (2003), "Nonparametric Model Calibration Estimation in Survey Sampling", *Journal of American Statistical Association*.,100, 1429-1442.

[8] Wu, C, & Sitter, R.R. (2001), "A Model Calibration Approach to Using Complete Auxiliary Information from Survey Data", *Journal of American Statistical Association,* 96, 185-93.